

21世纪全国高等院校财经管理系列实用规划教材



统计学理论与实务

主　编　王雪秋　董小刚





说　明

本书电子版仅提供给高校任课教师使用，如有任课教师需要本书课件或其他相关教学资料，请联系北京大学出版社客服，微信手机同号：15600139606，扫下面二维码可直接联系。

由于教材版权所限，仅限任课教师索取，谢谢！



21 世纪全国高等院校财经管理系列实用规划教材

统计学理论与实务

主　编　王雪秋　董小刚
副主编　王明明　赵天宇
参　编　张纯荣　赵　晖　杜春晶

北京大学出版社
PEKING UNIVERSITY PRESS

内 容 简 介

本书以统计学基本步骤为主线，从最为基础的收集和整理数据开始，渐次介绍数据分析的两种方法：描述性统计分析和推断性统计分析。对变量的分析也由研究单一变量间的关系逐步过渡到对二维变量及多元变量的分析，其间适时配以案例。考虑到学生的向量代数的基础，回归分析中主要是以单个方程为主。

本书共分 8 章，具体内容包括：总论、统计数据的收集与处理、统计数据的整理与图形展示、统计数据的指标度量、参数估计、假设检验、方差分析、相关与一元回归分析。

本书既可作为高等院校财经管理类专业的本、专科生教材，也可作为相关工作人员自学参考用书。

图书在版编目(CIP)数据

统计学理论与实务/王雪秋，董小刚主编. —北京：北京大学出版社，2015.8
(21 世纪全国高等院校财经管理系列实用规划教材)
ISBN 978-7-301-24467-8
Ⅰ. ①统… Ⅱ. ①王…②董… Ⅲ. ①统计学—高等学校—教材 Ⅳ. ①C8
中国版本图书馆 CIP 数据核字(2014)第 147742 号

书　　名　统计学理论与实务
著作责任者　王雪秋　董小刚　主编
责任编辑　王显超
标准书号　ISBN 978-7-301-24467-8
出版发行　北京大学出版社
地　　址　北京市海淀区成府路 205 号　100871
网　　址　http://www.pup.cn　新浪微博：@北京大学出版社
电子信箱　pup_6@163.com
电　　话　邮购部 62752015　发行部 62750672　编辑部 62750667
印 刷 者
经 销 者　新华书店
787 毫米×1092 毫米　16 开本　13.75 印张　312 千字
2015 年 8 月第 1 版　2015 年 8 月第 1 次印刷
定　　价　30.00 元

21世纪全国高等院校财经管理系列实用规划教材

专家编审委员会

丛　书　序

我国越来越多的高等院校设置了经济管理类学科专业，这是一个包括理论经济学、应用经济学、管理科学与工程、工商管理、公共管理、农林经济管理、图书馆、情报与档案管理 7 个一级学科门类和 31 个专业的庞大学科体系。2006 年教育部的数据表明，在全国普通高校中，经济类专业布点 1518 个，管理类专业布点 4328 个。其中除少量院校设置的经济管理专业偏重理论教学外，绝大部分属于应用型专业。经济管理类应用型专业主要着眼于培养社会主义国民经济发展所需要的德智体全面发展的高素质专门人才，要求既具有比较扎实的理论功底和良好的发展后劲，又具有较强的职业技能，并且又要求具有较好的创新精神和实践能力。

在当前开拓新型工业化道路，推进全面小康社会建设的新时期，进一步加强经济管理人才的培养，注重经济理论的系统化学习，特别是现代财经管理理论的学习，提高学生的专业理论素质和应用实践能力，培养出一大批高水平、高素质的经济管理人才，越来越成为提升我国经济竞争力、保证国民经济持续健康发展的重要前提。这就要求高等财经教育要更加注重依据国内外社会经济条件的变化，适时变革和调整教育目标和教学内容；要求经济管理学科专业更加注重应用、注重实践、注重规范、注重国际交流；要求经济管理学科专业与其他学科专业相互交融与协调发展；要求高等财经教育培养的人才具有更加丰富的社会知识和较强的人文素质及创新精神。要完成上述任务，各所高等院校需要进行深入的教学改革和创新，特别是要搞好有较高质量的教材的编写和创新工作。

出版社的领导和编辑通过对国内大学经济管理学科教材实际情况的调研，在与众多专家学者讨论的基础上，决定编写和出版一套面向经济管理学科专业的应用型系列教材，这是一项有利于促进高校教学改革发展的重要措施。

本系列教材是按照高等学校经济类和管理类学科本科专业规范、培养方案，以及课程教学大纲的要求，合理定位，由长期在教学第一线从事教学工作的教师编写，立足于 21 世纪经济管理类学科发展的需要，深入分析经济管理类专业本科学生现状及存在的问题，探索经济管理类专业本科学生综合素质培养的途径，以科学性、先进性、系统性和实用性为目标，其编写的特色主要体现在以下几个方面：

（1）关注经济管理学科发展的大背景，拓宽理论基础和专业知识，着眼于增强教学内容与实际的联系和应用性，突出创造能力和创新意识。

（2）体系完整、严密。系列涵盖经济类、管理类相关专业以及与经管相关的部分法律类课程，并把握相关课程之间的关系，整个系列丛书形成一套完整、严密的知识结构体系。

（3）内容新颖。借鉴国外最新的教材，融会当前有关经济管理学科的最新理论和实践经验，用最新知识充实教材内容。

（4）合作交流的成果。本系列教材是由全国上百所高校教师共同编写而成，在相互进行学术交流、经验借鉴、取长补短、集思广益的基础上，形成编写大纲。最终融合了各地特点，具有较强的适应性。

（5）案例教学。教材融入了大量案例研究分析内容，让学生在学习过程中理论联系实际，特别列举了我国经济管理工作中的大量实际案例，这可大大增强学生的实际操作能力。

（6）注重能力培养。力求做到不断强化自我学习能力、思维能力、创造性解决问题的能力以及不断自我更新知识的能力，促进学生向着富有鲜明个性的方向发展。

作为高要求，经济管理类教材应在基本理论上做到以马克思主义为指导，结合我国财经工作的新实践，充分汲取中华民族优秀文化和西方科学管理思想，形成具有中国特色的创新教材。这一目标不可能一蹴而就，需要作者通过长期艰苦的学术劳动和不断地进行教材内容的更新才能达成。我希望这一系列教材的编写，将是我国拥有较高质量的高校财经管理学科应用型教材建设工程的新尝试和新起点。

我要感谢参加本系列教材编写和审稿的各位老师所付出的大量卓有成效的辛勤劳动。由于编写时间紧、相互协调难度大等原因，本系列教材肯定还存在一些不足和错漏。我相信，在各位老师的关心和帮助下，本系列教材一定能不断地改进和完善，并在我国大学经济管理类学科专业的教学改革和课程体系建设中起到应有的促进作用。

刘诗白

刘诗白　现任西南财经大学名誉校长、教授，博士生导师，四川省社会科学联合会主席，《经济学家》杂志主编，全国高等财经院校《资本论》研究会会长，学术团体“新知研究院”院长。

目　录

第 1 章 总 论

教学目标

1. 掌握统计学的含义及研究对象。
2. 掌握统计学的分类。
3. 掌握统计数据的类型。
4. 掌握统计学中的几个基本概念。

1.1 统计学概述

日常生活中，人们经常会使用“统计”这一专业术语，同时也会在有关的媒体中经常看见一些资料使用统计数据、图表等形式。本章将详细讲解统计学的基本原理，包括统计及统计学的含义、统计学研究的对象及特点、统计学的应用领域、统计学的分类、统计数据的类型及统计学中几个基本的概念。

1.1.1 统计及统计学的含义

统计指对某一现象有关数据的收集、整理、计算和分析等的活动。在统计的应用中，人们对“统计”一词的理解一般有 3 种含义：统计工作、统计资料和统计学。

定义 1.1 统计工作是指利用科学的方法收集资料、整理资料、分析资料和提供关于社会经济现象数量关系的工作总称。

统计工作是统计资料和统计学的基础。

定义 1.2 统计资料是指通过上面所说的统计工作取得的、用来反映社会经济现象的数据资料总称。

统计工作的成果是统计资料，统计工作所取得的各项数字资料及有关文字资料，通常反映在统计表、统计图、统计手册、统计年鉴、统计资料汇编和统计分析报告中。

定义 1.3 统计学是指研究如何收集数据、如何整理数据、如何分析数据和最后解释数据，并从数据中得出规律的一门科学。

统计学既是统计工作经验的理论概括，又是指导统计工作的原理、原则和方法。

统计工作、统计资料、统计学三者之间的关系：利用统计学的理论，指导统计工作，

最后得出统计资料。其中统计学是关于数据的科学，它提供了有关数据收集、数据处理、数据分析、数据解释并从数据中得出结论的方法。

1.1.2 统计学研究的对象及特点

1. 统计学研究的对象

统计学研究的内容是数据，而数据是社会经济现象的数量特征和数量关系的表现，从数据中找出经济规律性。因此，统计学的研究对象为大量社会经济现象的数量特征和数量关系，以揭示其规律性。

2. 统计学研究对象的特点

1) 总体性

一般情况下，统计学研究对象是社会经济现象总体或自然现象总体的数量特征，研究的方法是利用样本的信息推导出总体的数量特征。

例如，要研究某地区居民的收入水平，目的不在于了解个别居民的收入状况，而是要通过对很多个别居民收入状况的了解，达到对全区居民总体收入水平的认识。再如，某企业对其生产的一批日光灯管的平均使用寿命进行研究，不可能把这批日光灯管全部点亮来研究其平均寿命，因为该种研究属于破坏性研究，只能从这批日光灯管中抽取一组样本，对样本进行实验，通过样本的平均寿命来得知总体的平均寿命。

2) 数量性

统计学研究对象的数量性。

对象的数量性包含 3 个方面的内容：研究对象数量的多少、研究现象间的数量关系和研究对象的质与量间的关系。即通过各种不同的统计指标和指标体系来反映研究对象总体的规模、水平、速度、比例、效益和趋势等。

例如，我国的人口数量构成及其发展趋势、人口结构的构成及发展趋势、国民生产总值的总量构成等。再如，某投资者筹备一超市，筹备前就要研究选址是否会影响超市的营业利润。

1.1.3 统计学的应用领域

统计学的研究内容为数据，只要有数据的存在就会用到统计学。随着定量研究重要性的提高，现代统计学的应用日益广泛，理、工、农、医、文、经，都要用到统计学的。例如，政府部门利用统计学进行宏观调控和管理；日常生活中，统计学是企业管理与决策的依据等。下面主要介绍统计学在经济管理中的一些应用。

1. 企业开发新产品

企业为了在激烈的市场竞争中求得平稳的发展，必然要根据市场的变化在适当的时候引入新产品。引入新产品前，企业要对新产品进行市场定位，包括市场分析、对新产品的价格进行市场定位、对消费群体进行定位等，只有全面分析，新产品开发才能获胜。同时，新产品进入市场后，企业要不断进行产品跟踪调查，不断完善新产品。而这些离不开统计学，它们需要统计学提供可靠的数据，对数据进行分析，得出决策的信息。

2. 财务分析

上市公司的财务报表中的数据是投资者进行决策的重要参考依据。投资者分析上市公司的财务报表数据的定量关系，最后进行决策。企业自身的投资也离不开财务数据的分析。

3. 竞争对手的研究

随着社会经济的发展，各个行业内部之间的竞争也日益加剧。企业要想取得优势，抢占时机，就要不断地了解竞争对手，摸清对手的竞争策略，发现其弱势，利用自己的优势，制定获胜的策略，而做这些的前提是要进行数据分析。

1.2 统计学的分类

1.2.1 描述统计学和推断统计学

1. 描述统计学和推断统计学的定义

统计学按照统计数据分析的方法不同，分为描述统计学和推断统计学。

定义 1.4　描述统计是指研究数据收集、处理数据和描述数据的统计学分支。

描述统计学的内容包括如何取得要研究的数据、用什么样的图表对数据进行处理和显示，进而通过综合、概括与分析，得出反映所研究现象的一般性特征。

例如，反映长春大学光华学院经济系某年级的统计学成绩的条形图即属于描述统计学，其统计学成绩数据见表 1-1 所示。

表 1-1　统计学成绩

成绩结果	频　数
优秀	7
良好	15
中等	16
及格	14
不及格	3
合计	55

该组数据采用条形图来描述，利用 Excel 软件来绘制，绘制结果如图 1.1 所示。

定义 1.5　推断统计学是指通过研究如何根据统计样本的资料计算样本的特征信息，来推断总体相关的特征的方法。

例如，从一个果园中采摘 40 个橘子，利用这 40 个橘子的平均重量来估计整个果园所有橘子的平均重量。

2. 描述统计学与推断统计学的关系

一般来讲，描述统计学是现代统计学的前期工作，是推断统计学的基础；推断统计学是现代统计学的核心，是描述统计学的发展。这两部分是统计学的两个不可缺的组成内容。

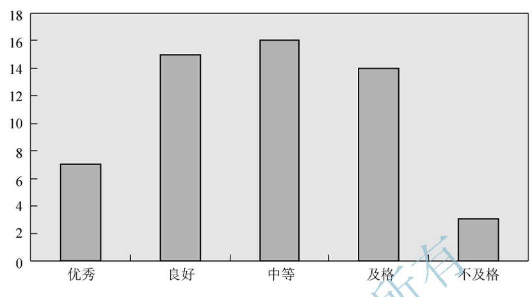


图 1.1　统计学成绩条形图

例如，想研究某城市居民的消费水平，而往往这个城市的所有居民消费水平的数据是很难收集的，这时，需要对这个城市的一小部分居民的消费状况数据进行收集，通过描述统计学，了解样本居民的消费状况，而后用到推断统计学，在对样本居民消费状况的了解情况下，达到对全市居民总体消费水平的认识。

1.2.2　理论统计学和应用统计学

1. 理论统计学和应用统计学的定义

统计学按照研究的目的不同，可分为理论统计学和应用统计学。

理论统计学把研究对象一般化、抽象化，以概率论为基础，从纯理论的角度，对统计方法加以推导论证，以归纳方法研究随机变量的一般规律。

应用统计学侧重于统计学的应用，它研究如何应用统计学理论和方法，对实际的问题进行研究，以揭示其各种经济现象的规律性。其目的是解决经济存在的问题，对经济增长进行预测。

2. 理论统计学和应用统计学的关系

理论统计学和应用统计学的关系十分密切。理论统计学为应用统计学提供了统计的理论和方法；而应用统计学是理论统计学的发展或延伸。

1.3　统计学的基本内容

从统计学的定义中可以看出统计学的核心是数据，所以在学习统计学内容之前要掌握数据的类型。

1.3.1　统计数据的类型

统计学在研究事物的数量方面是离不开数据的，如研究股票价格变动水平就要收集数据，计算出股票价格指数的指标来分析等，但数据不同，分析的方法不同。

1. 分类数据、顺序数据、数值型数据

按照计量尺度的不同，统计数据可分为分类数据、顺序数据和数值数据。

定义 1.6　只归于某一类别的非数字型数据，称为分类数据。

分类数据是对事物进行分类，该数据表现的是文字叙述，如人口按性别可分为男、女两类。

定义 1.7　只能归于某一有序类别的非数字型数据，称为顺序数据。

顺序数据也是对事物进行分类的结果，如某大学的选修课以优、良、中、及格和不及格分成 5 级。

定义 1.8　按数字尺度测量的观察值，称为数值数据。

数值数据是对事物进行了准确的测量，表现具体的数值，如某大学的统计学成绩。

在这里要注意以下两点。

(1) 分类数据是指归于某一类别非数字型数据，而顺序数据是归于某一有序的类别，即在分类数据的基础上加一个条件“有序”。从这里可以得出，顺序数据比分类数据高一层次。

(2) 数值数据是对事物进行了准确的测量，表现具体的数值。例如，某专业学生的“英语成绩”为数值数据，但有时为了了解成绩的状况，要对数据进行分析，即 90 分以上的学生归于优秀，80～89 分的学生归于良好，70～79 分的学生归于中等，60～69 分的学生归于及格，最后 60 分以下的归于不及格。所以说，从这个例子可以得知 3 种数据，最高级数据为数值数据，其次是顺序数据，最低级的数据为分类数据。

2. 观测数据和实验数据

按照统计数据的收集方法不同，统计数据可分为观察数据和实验数据。

定义 1.9　通过调查或者观测而收集到的数据，称为观察数据。

观察数据是在没有对事物进行人为控制的条件下得到的数据，有关社会经济现象的统计数据几乎都是观察数据。

例如，某一时间段，某市交通广播电台发布信息：延安大街由南向北车流量比较大，请各位司机避让。此数据“车流量大”就是观察数据。再如，某人对寄居在家中的一窝燕子很感兴趣，他每天观察并记录燕子飞出飞入的时间，以及它们喂养小燕子的习惯，那么他得到的燕子飞出飞入的时间数据即为观察数据。

定义 1.10　在实验中控制实验对象而收集到的数据，称为实验数据。

实验数据是对事物进行了人为控制而收集到的数据。在自然科学领域中所使用的统计数据大多是实验数据。

例如，某医药企业研发新药，数据通常是通过对小白鼠进行实验而得到的实验数据。

3. 截面数据和时间序列数据

按照被描述的对象与时间之间的关系，统计数据分为截面数据和时间序列数据。

定义 1.11　在相同或近似相同的时间点上所收集的数据，描述现象在某一时点的变化情况，称为截面数据。

定义 1.12　在不同时间上所收集的数据，用来描述现象随时间而变化情况的数据，称为时间序列数据。

例如，1994—1999 年我国国内生产总值见表 1-2 所示。

表 1-2　1994—1999 年我国国内生产总值

单位：亿元人民币

年　份	第一产业	第二产业	第三产业
1994	9 457.2	22 372.2	14 930
1995	11 993	28 537.9	17 947.2
1996	13 844.2	33 612.9	20 427.5
1997	14 211.2	37 222.7	23 028.7
1998	14 552.4	38 619.3	25 173.5
1999	14 457.2	40 417.9	27 035.8

其中行数据为截面数据，列数据为时间序列数据，即有 3 组时间序列数据，6 组截面数据。

1.3.2　统计学中的几个基本概念

统计学主要是想从样本的信息推断出总体的特征值。在进行研究时，涉及很多概念，其中有几个概念是经常用到的。

1. 总体和样本

定义 1.13　包含所研究的全部个体(数据)的集合，称为总体。

例如，要检验一批奶粉是否合格，那么这批奶粉构成的集合就是总体，其中每一袋奶粉就是总体中的个体。

总体根据包含的个体数目是否可数，分为有限总体和无限总体，即总体的范围确定有时比较容易，有时较难。例如，要检验一批灯泡的使用寿命，那么这批灯泡构成的集合就是总体，较易确定。再如，某企业推出一种新产品，想了解消费者是否喜欢，这时它就需要先确定消费对象，即要确定消费这种新产品的消费者总体，这时该总体的范围确定就较难。所以说，在实务中，我们可根据自己的研究目的来定义总体。

定义 1.14　从总体中抽取一部分个体重新组成一个集合，称为样本。

从总体中抽取一部分个体作为样本，目的是要根据样本提供的有关信息去推断总体的信息。

例如，估计一批灯泡的平均寿命，不能把所有灯泡点亮，因为这种试验属于破坏性试验。所以要从这批灯泡中随机抽取一小部分作为样本，进行测试，从而得出样本的平均寿命，通过此信息推断出这批灯泡的平均寿命。

再如，研究全国大学生平均月消费结构的状况。通常全国大学生这个总体的数据不易得到，只能抽取样本，通过对样本的信息研究推断出总体的信息。

定义 1.15　构成样本的个体数目，称为样本容量，或称为样本量。

2. 参数和统计量

定义 1.16　用来描述总体特征的概括性数字度量，称为参数。

参数是研究者想要了解总体信息的特征值。在统计学中，我们最关心的是总体的均值、

总体的方差、总体的比例等。这些参数通常是未知的。

例如，某企业想了解今天生产的这批灯泡的平均寿命，即总体的均值；某投资企业想了解它的投资组合的风险，即总体方差；某企业想了解今天生产的这批灯泡的次品率，即总体的比例。

在统计中，总体参数通常用希腊字母表示，如总体平均数用 μ 表示，总体标准差用 σ 表示，总体比例用 π 表示等。

定义 1.17　用来描述样本特征的概括性数字度量，称为统计量。

统计量是根据样本数据计算出来的一个已知的量。通常，最关心的是样本的平均数、样本的方差、样本的比例等。样本统计量也用英文字母表示。例如，样本平均数用 $\bar{x}$ 表示，样本标准差用 s 表示，样本比例用 p 表示等。

抽取样本的目的是根据样本的统计量估计总体的参数，即用样本平均数 $\bar{x}$ 估计总体平均数 μ，用样本标准差 s 估计总体的标准差，用样本比例 p 估计总体比例 π。

3. 变量

定义 1.18　说明现象某种特征的概念，称为变量。

变量按照不同的划分标准有不同的分类：

(1) 按数据的类型，变量可分为分类的变量、顺序变量和数值变量。

(2) 按数值是否可数，变量分为离散型变量和连续型变量。

1.4　案例分析：啤酒市场的调查与分析

随着经济的快速发展，人们的生活水平也日益提高了。在经济不发达的时候，人们对一种产品是否消费，主要是受产品的价格因素影响，但随着经济的发展，人们的生活水平也在不断地提高，人们对一种产品的消费，不单单会受产品的价格因素影响，还有其他因素，如品牌、产品的售后服务、产品的外观、性能等因素的影响。

啤酒已成为一种日常消费品进入千家万户，且啤酒市场的竞争也一直从未停止过，各类啤酒犹如雨后春笋般不断地推陈出新。2003 年，自我国加入 WTO(World Trade Organization，世界贸易组织)后，开放的中国市场已逐渐融入到了世界经济的均衡游戏中。尤其从 2005 年开始，外资进入中国啤酒业的步伐更为快速，随着外资收购速度的加快，中国啤酒市场的竞争正在发生变化，要想在这激烈的竞争市场中处于不败之地，抓住市场、扩展市场才是唯一的出路。所以有必要进行新一轮的啤酒市场调查与分析，以准确地知道是哪些因素影响消费者的购买行为。

要想准确地知道是哪些因素影响消费者的购买行为，首先要收集数据，其次要整理数据，再次要分析数据，最后解释数据，从数据中得出想要的信息。

收集数据前要确定研究的总体。从上面的分析可知，总体是购买啤酒的消费者，这一总体的范围很难确定，即总体的数据很难收集，所以使用推断性统计得出总体的信息，即收集样本的数据，根据样本的信息来推断总体的信息。

啤酒市场的调查与分析流程如图 1.2 所示。

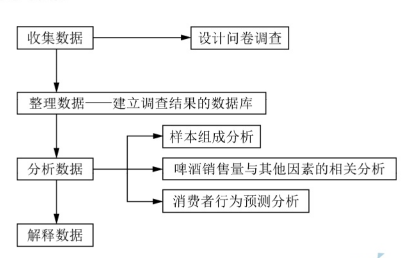


图 1.2 啤酒市场的调查与分析流程

习 题

1．指出下列数据的类型。

(1) 年龄。

(2) 工资。

(3) 汽车产量。

(4) 性别。

(5) 购买商品时的支付方式(现金、信用卡、支票)。

(6) 企业类型。

(7) 员工对企业某项改革措施的态度(非常赞成、赞成、中立、反对、非常反对)。

2．一家研究机构从 IT 从业者中随机抽取了 200 人作为样本进行调查，其中 40%的人回答他们的月收入在 4 000 元以上，60%的人回答他们的消费支付方式是现金。试问：

(1) 该研究的总体是什么？样本是什么？样本容量为多少？

(2) 月收入是什么类型数据？消费支付方式是什么类型数据？

3．判断下列哪些是推断统计，哪些是描述性统计。

(1) 从一个果园中采摘 40 个苹果，利用这 40 个苹果的平均重量估计果园的所有苹果的平均重量。

(2) 用条形图描述某大学某专业学生的成绩状况。

(3) 调查一个城市拥有汽车家庭的比例，估计全国拥有汽车家庭的比例。

4．为了估计某城市中拥有汽车的家庭比例，抽取了样本容量为 500 个家庭的样本，得到拥有汽车家庭的比例为 35%。根据这一信息估计这个城市拥有汽车家庭的比例为 32%。试问，哪个数据是参数？哪个数据是统计量？

第 2 章　统计数据的收集与处理

教学目标

1. 了解数据收集的主要方式和方法，以及各种方法的特性。
2. 了解统计调查方案的设计。
3. 掌握统计调查问卷设计的主要内容。
4. 掌握数据收集的软件操作过程。

导入案例

2010 年第六次全国人口普查

列宁曾说过“统计不是把数字随便填列到几个格子里去，而是应当用数字来说明所研究的现象在实际生活中已经充分呈现出来或正在呈现出来的各种社会类型”。

第五次全国人口普查查清了我国人口在数量、结构、地区分布、受教育程度、迁移流动和居住环境等方面的变化情况，为科学制定国民经济和社会发展规划，统筹安提排人民的物质和文化生活，实施可持续发展战略，构建社会主义和谐社会，提供科学准确的统计信息支持。我国于 2010 年 11 月 1 日 00:00 开始进行了第六次全国人口普查。通过人口普查的结果，可以查清全国农民工问题、人口老龄化的进展状况等，为研究制定“十二五”规划提供依据，并为社会公众提供人口信息服务。

根据实际研究的问题，进行统计分析，而任何应用统计分析问题首先要取得数据，即收集数据是统计分析的前提，同时收集到可靠准确、高质量的数据是统计研究的重要内容之一。本章主要介绍数据的收集方法，使学生掌握取得数据的方法。具体内容包括数据的来源及不同来源的数据的处理方法，重点介绍直接来源数据的收集方法及处理的过程。

2.1　统计数据的来源

从使用者的角度看，统计数据的来源主要有两种：一是间接来源，间接来源数据又称二手数据，即别人已收集到的数据，不用研究人员去自己收集；二是直接来源，直接来源数据又称一手数据，即别人没有收集过的数据，需要研究人员自己去收集数据。

2.1.1 统计数据的间接来源与处理

1. 间接来源数据

如果与研究内容有关的原信息已存在，人们只需对这些原始信息进行重新加工、整理以满足所需的统计数据，则称这种数据为间接来源的数据。例如，对改革开放以来吉林省区域经济发展趋势的研究。针对研究内容要收集改革开放以来吉林省每年的评价区域经济发展的指标国内生产总值，而这些数据是已存在的，即只需到统计年鉴中找到数据，重新整理，这样的数据就属于间接来源数据。

通常状况下，间接数据主要来源于社会经济统计部门公开出版或公开报道的各种报刊书籍，如公开出版的《中国统计年鉴》《中国社会统计年鉴》等，当然也有些是尚未公开的数据。

2. 间接来源数据的处理

相对于直接来源数据，间接来源数据有很多优点，如收集起来比较容易，收集数据的成本较低，用时也较短。

虽然间接来源数据对于使用者来说既经济又方便，但使用时需要使用者保持谨慎的态度，因为间接来源数据并不是针对研究人员的研究内容而收集的数据，所以这种数据可能是有欠缺的，如资料的相关性不够，口径可能不一致，数据可能不准确，失去时效性等。因此，在使用间接来源数据时需要进行评估，一般情况下，评估间接来源数据时考虑以下一些因素：

(1) 数据是由谁搜集的。主要考察数据搜集者的实力和社会信誉度。例如，对于全国性的宏观数据，与某个专业性的调查机构相比，政府有关部门公布的数据可信度更高。

(2) 数据收集的目的。为了某个集团利益而搜集的数据是值得怀疑的。

(3) 数据的搜集方法。数据的搜集方法有很多，不同的搜集方法所得到的数据的解释力和说服力是不同的。如果不了解搜集数据的方法，很难对数据的质量做出客观的评价，即数据的质量来源于数据的产生过程。

(4) 间接来源数据的搜集时间，即注意数据的时效性。因为使用“过时”数据会影响研究内容的结果。

(5) 间接来源数据的一致性。主要表现为数据的计算口径是否相同。例如：评价几家保险公司本月健康险的赔付，需要搜集这几家保险公司的健康险的本月赔付数据进行比较，这时需要注意这几家保险公司的赔付数据的计算口径是否一致，即赔付数据是否包括了已支付、已发生或已报告赔款。

最后，在引用间接来源数据时，应注明数据的来源，以尊重他人的劳动成果。

2.1.2 统计数据的直接来源与处理

当间接来源数据(二手数据)无法满足需要的时候，可以亲自搜集数据，这种数据称为直接来源数据，又称一手数据。直接来源数据的主要来源有两个。一是通过专门组织的统计调查获得数据，即调查数据。统计调查是取得社会经济数据的重要手段。二是通过科学试验得到的数据，即实验数据。科学试验是取得自然科学数据的主要手段。其中调查数据有两种常用的搜集方式——普查和抽样调查。普查是指对总体中的个体进行逐一调查。抽

样调查是按照一定的筛选规则选择一部分个体进行调查。这里主要介绍抽样调查数据。

1. 普查

普查是专门组织的，一般用来全面调查属于一定时点上的社会经济现象的数量。例如，世界各国一般都定期地进行各种普查，以便掌握有关国情、国力的基本统计数据。

普查的目的是掌握特定社会经济现象的基本全貌，如为国家制定有关政策或措施提供依据等。我国进行的普查主要有人口普查、农业普查和经济普查。我国人口普查是每隔 10 年进行一次，每逢年份的末尾为“0”的年份进行人口普查，如在 2010 年 11 月 1 日零时进行；我国农业普查也是 10 年进行一次，每逢年份的末尾为“6”的年份进行农业普查；经济普查是 5 年进行一次，每逢年份的末尾为“3”和“8”的年份进行。

由于普查是对研究对象进行全面的调查，即普查涉及的面广，需要大量的人力、物力、财力和时间，因此普查间隔的时间较长。而对于微观经济主体来说，需要时时刻刻了解外境及内境的变化情况，需要时时刻刻进行调查，所以往往会采用抽样调查。

2. 抽样调查

抽样调查是指从总体中随机抽取一部分单位(个体)作为样本进行调查，并根据样本调查结果来推断总体特征的数据采集方法。抽样调查是实际中应用最广泛的一种调查方式。抽样调查具有以下特点。

(1) 经济性。调查的样本单位通常是总体单位的一小部分，调查的工作量小，因而可以省力、省时，调查的成本较低。

(2) 时效性较高。由于抽样调查只是调查总体中的一小部分，因此调查的准备时间较短，调查时间也较短，数据的处理时间同样较短，从而提高了数据的时效性。

(3) 适用面广。从适用范围和问题来看，抽样调查既能调查到全面调查所能调查的现象，同时也能调查到全面调查调查不到的现象。

(4) 准确性较高。抽样调查中的样本误差，在调查前就可以根据调查样本数量和总体中各单位之间的差异程度进行计算，可以把样本误差控制在一定范围之内，调查结果的准确程度比较有把握。

2.2 抽样调查数据的收集

普查工作耗费人力、物力较大，时间较长，常因总体资料取得不易而无法进行，而抽样调查的方法因具有准确性高、成本低、速度快、应用面广等优点，是企业中运用最为普遍的调查方式，也是市场经济国家在实地调查方法上的必然选择。在一项调查工作开始前，首先要对调查工作进行设计，即设计调查方案。

2.2.1 调查方案的设计

调查方案的设计是指事先制订出一个科学、严密、可行的工作计划并组织实施，以便在调查过程中统一认识、统一内容、统一方法、统一步调，圆满完成调查任务，即包括整个调查工作过程的内容。

调查方案的设计包括以下内容。

1. 确定调查目的

明确调查目的是调查设计的首要问题，只有明确了调查的目的，才能确定研究人员的调查范围和内容及方法。

例如，公司决定开发一款新产品，在具体的研发之前，想要了解目前市场上同类产品的销售情况、价位、在消费者心目中的印象及消费者可接受的产品价位等信息，此时市场调查的工作便是通过对消费者和竞争产品的调查分析，找准新产品的目标市场和目标消费群体，即此时的问卷调查的问题要针对这一目的而展开。

2. 确定调查对象和调查单位

在明确调查目的之后，针对目的，要明确需要什么样的数据，明确数据如何得到，明确数据由谁提供等，即要明确调查对象和调查单位。

在确定调查对象和调查单位时，需要注意确定的方法必须用科学的理论作指导，严格规定调查对象的含义，并划分它与其他有关对象的界限，以免调查记录由于界限不清而发生差错。

例如，以城市职工为调查对象，就应明确城市职工的含义，划清城市职工与非城市职工、城市职工与城市居民等概念的界限。

3. 决定抽样调查的方法及数据收集的方法

1) 抽样调查的方法

抽样调查的方式有很多，可以将这些不同的方式分为两类：概率抽样和非概率抽样。

(1) 概率抽样。概率抽样也称随机抽样，是指按照随机原则进行的抽样，总体中每个单位都有一定的机会被选入样本。其中随机原则是指在抽取样本时排除主观上有意识地抽取调查单位，使每个单位都有一定的机会被抽中。注意：随机不是随便，随机用概率来描述，而随便则带有人为主观的因素。这里指的每个单位都有一定的机会被选入样本，不等同于每个单位都以相同的概率被抽中，而是指每个单位被抽中的概率为非零。

概率抽样的方式有以下 5 种。

① 简单随机抽样。简单随机抽样也称纯随机抽样，就是在总体单位中不进行任何有目的的选择，完全按随机原则抽取样本单位。由于市场调研的总体范围较广，总体内部各个个体之间的差异较大，一般不直接采用这种抽样的方法，而是与其他抽样方法相结合。

② 分层抽样。分层抽样是将抽样单位按某种特征或某种规则划分为不同的层，然后从不同的层中独立、随机地抽取样本。将各层的样本结合起来，对总体的目标量进行估计。

例如，研究在校大学生对图书馆的利用率。抽取样本时，先按专业将所有的在校大学生分不同层，然后按照总体中不同层的学生比例，从不同专业中抽取若干名学生组成在校大学生的抽取样本。

分层抽样的优点有很多，如抽样方法保证了样本中包含有各种特征的抽样单位，样本的结构与总体的结构比较相近，从而可以有效地提高估计的精度；再如，分层抽样不仅可以对总体参数进行估计，还可以对各层的目标量进行估计。

③ 整群抽样。整群抽样是指先将总体中若干个单位合并为组(群)，抽样时直接抽取群，然后对选中的群中的所有单位进行全部调查。

例如，研究在校大学生对统计学这门课程的认识。先采用整群抽样，按学校的名称把在校大学生分成若干个群，如清华大学的学生、北京大学的学生……然后从这些群中抽取一个群，作为总体的样本，对抽中的群中所有在校大学生实施调查。

整群抽样的优点：群通常由那些地理位置邻近的或隶属于同一系统的单位构成，因此调查的地点相对集中，从而节省了调查费用，方便了调查的实施。

整群抽样的缺点：估计的精度较差，因为同一群内的单位或多或少地有些相似，在样本量相同的条件下，抽样误差通常较大，导致精度较差。

④ 系统抽样。系统抽样是指将总体中的所有单位按一定的顺序排列，在规定的范围内随机地抽取一个单位作为初始单位，然后按事先规定好的规则确定其他样本单位。

例如，先将总体中的个体随机排序并编号，规定抽取号码为双号，随机地抽取 4 号作为初始单位，共抽取样本容量为 40 的样本。按照系统抽样抽取的样本中个体有 4 号、6 号、8 号……直到抽取 40 个个体为止。

系统抽样的主要优点是操作简单，如果对总体内的单位进行有组织的排列，可以有效地提高估计的精度。系统抽样的主要缺点是对估计量方差的估计比较困难。

⑤ 多阶段抽样。多阶段抽样是指分两个及两个以上的阶段从总体中抽取样本的一种抽样调查方法，即先粗分，再细分，然后再微分。

例如，对“实施全省性的防治犯罪相关问题”进行民意调查，决定采用多阶段的抽样方法。

首先，针对全省居住人口，按犯罪率高低从各市、县、区依一定比例随机抽出 100 个个体(各市、县、区均有)。

其次，在这 100 个个体中，以镇、街道为类，在同一个体中抽出 3 个村(居委会)。

最后，在村(居委会)以户为单位，随机抽出 5 户作为样本，所以最后样 3 数为 1 500。

以上介绍了几种常见的概率抽样方式，主要的优点是，可以依据调查结果，计算估计量误差，从而得到对总体目标量进行推断的可靠程度。也可以按照要求的精确度，计算必要的样本单位数目，此类问题将在第 5 章具体介绍。

(2) 非概率抽样。非概率抽样是指抽样时不是依据随机原则，而是根据研究目的对数据的要求，采用某种方式从总体中抽出部分单位对其进行调查研究。非概率抽样有以下几种常见的方法。

① 方便抽样。方便抽样又称偶遇抽样，即调查员依据方便的原则，自行确定抽取的样本单位。

例如，调研者在路上或其他地方，如快餐店或便利店等，拦下行人进行访问就是一种方便抽样。再如，研究某城市居民购房需求的状况，调研者在此城市房交会门口拦下每个从房交会出来的人进行调查研究。

方便抽样的优点：简便易行，能及时获得所需要的信息数据；省时省力，节省调研经费；效率很高，并能为非正式的探索性研究提供很好的数据源。

方便抽样的缺点：样本的偶然性较高，存在选择的偏差，即样本的代表性较差，调查的结果可信度较低。所以一般情况下，此抽样方法只适用于探索性的调查或正式调查前的预调查。

② 判断抽样。判断抽样是指研究人员根据自己的经验、判断和对研究对象的了解，有目的地选择一些个体作为样本。判断抽样一般情况有两种做法。

一种是由专家判断决定所选样本，即选择最能代表普遍情况的群体作为样本，其中普遍情况的群体一般选择“多数型”或“平均型”为样本进行调查研究。

多数型是指选择的样本在调研的总体中占多数的单位。例如，调查中国钢铁行业的管理机制、运营机制及改革等状况，所挑选的样本单位一定得避开鞍钢、宝钢和首钢等几家大型企业，原因是它们的钢铁产量占全国钢铁产量的大半，但是它们的管理水平、运营能力等不能代表众多钢铁企业的现状。

平均型是指选择的样本是调研总体中的能代表平均水平的单位。例如，某企业要调查其自身产品与竞争对手产品的销售情况，根据主观判断选择了一些同时对销售双方产品有影响的、非常有代表性的零售商店作为样本。

另一种是利用统计判断选取样本，即利用总体的全面统计资料，按照主观设定的某一标准的样本。例如，调查我国钢铁的产量状况，这时只需对我国的鞍钢、宝钢和首钢等几家大型企业进行调查，原因是它们的钢铁产量占全国钢铁产量的大半，对这几家进行了解就相当于掌握了总体产量的状况。

判断抽样的方法成本较低，也容易操作，但由于样本是人为确定的，没有依据随机的原则，因而调查结果不能用于对总体有关参数进行估计。

③ 自愿样本。自愿样本指被调查者自愿参加，成为样本中的一个体，向调查人员提供有关信息，如参与报刊上和互联网上刊登的调查问卷活动，都属于自愿样本。

自愿样本的样本组成往往集中于某类特定的人群，即集中于对该调查活动感兴趣的人群，因此，这种样本是有偏的。但自愿样本仍可以给研究人员提供许多有价值的信息，它可以反映某类群体的一般看法。

④ 配额抽样。配额抽样是指随意选择被调查的个体，但在性别、年龄和社会阶层等方面有名额的限制。这种方法在商业调查中广泛使用，但抽样人群依赖于调查者的喜好和调查地点。

配额抽样的方法操作比较简单，而且可以保证总体中不同类别的单位都能包括在所抽的样本之中，使得样本的结构和总体的结构类似。

⑤ 滚雪球抽样。滚雪球抽样又称链式抽样，是指利用随机方法选出初始受访者，然后从初始受访者所提供的信息中取得新的具有某一特征的再次受访者，依次如此，最后通过少量的样本单位逐步获得较多的样本单位的方法。

滚雪球抽样的主要优点是容易找到那些属于特定群体的被调查者，调查的成本也较低。它适用于对特定群体进行研究的资料搜集。

2) 数据收集的方法

样本单位确定之后，对这些单位实施调查，即从样本单位得到所需要的数据。数据收集的方法主要有以下几种。

(1) 自填式。自填式指在没有调查员协助的情况下由被调查者自己填写，完成调查问卷。把问卷递送给被调查者的方法有很多，如调查员分发、调查员邮寄、调查员通过网络发送等，本书的案例以调查员邮寄方式为主。

(2) 面访式。面访式是指现场调查中调查员与被调查者面对面，调查员提问、被调查者回答的调查方式。这种调查方式回收率较高，但成本也较高。

(3) 电话式。电话式是指调查人员通过打电话的方式向被调查者实施调查。这种调查

的方式速度快，能够在短时间内完成调查。虽然电话式可以在短时间内得到数据，但它也有很多局限性，如对方较忙无时间接听，或无人接听，或被研究对象无电话等。

2.2.2 问卷调查的设计

问卷调查法又称问卷法，是以问题的形式设计问卷，问题体现要调查的内容，以统一的方式向被选取的样本实施调查，收集调查的内容。

设计问卷是调查的关键部分。因为设计问卷是调查者得到数据的方法，而数据的质量关系到最后的分析结果。所以说，完美的问卷必须具备两个功能：一是能将问题传达给被问的人；二是使被问者乐于回答，使调查者收集到有效的数据。要完成这两个功能，设计问卷时应当遵循一定的原则和程序，运用一定的技巧。

1. 问卷设计的原则

1) 有明确的主题

明确的主题是指调查的主要目标，即问卷中的问题要根据调查目标，从实际出发拟题，重点突出，不能出现可有可无的问题。

2) 问卷的结构要合理、逻辑性要强

问卷中的问题排列要合理，要符合应答者的思维程序。通常是先浅后深、先简后繁、先具体后抽象。这样也有助于调查者得到质量高的数据。

3) 通俗易懂

问卷中问卷要通俗易懂，不能使用较强的专业术语，要使应答者一目了然，并愿意如实回答。要达到此目的，需注意以下几点。

(1) 问卷中语气要亲切，符合应答者的理解能力和认识能力，避免使用专业用语。例如，“您认为软饮料的分销充分吗？”这个问题中引用了技术用语“分销”很难让所有的被调查者理解这个词的含义。

(2) 对敏感性问题要采取一定的技巧调查。例如，问“××牌的产品质优价廉，您是否准备选购”这样的问题将容易使被调查者因引导性提问得出肯定性的结论或因反感此种问法简单得出结论，这样不能反映消费者对商品的真实态度和真正的购买意愿，所以产生的结论也缺乏客观性，可信度偏低。

(3) 避免隐含选择、隐含假设。例如：“您目前从事什么事业？”这个问题就隐含了一个假设，假设所有的被调查者都有工作。

4) 控制问卷的长度

问卷的长度通常用回答问卷的时间来控制，时间在20分钟左右。也就是说，问卷设计不能浪费一个问句，也不要遗漏一个问句，即问题应简明扼要，应尽量避免太长的题目。

5) 便于资料的校验、整理和统计

问卷是抽样调查收集数据的主要方法，只有把数据整理后，才能进行数据分析，所以，设计问卷时一定要考虑资料的校验、数据的整理和统计。

2. 问卷设计的程序

问卷设计的程序包括下列几个步骤。

1) 把握调研的目的和内容

问卷设计的第一步就是要把握调研的目的和内容。这一步骤的实质是规定设计问卷所需的信息。

2) 确定调查方法的类型

不同类型的调查方式对问卷设计是有影响的。

例如，在面访调查中，如果是入户访问的话，被调查者可以看到问题并可以与调查人员面对面地交谈，因此可以询问较长的、复杂的和各种类型的问题。如果是街上进行拦截式的面对面访谈就比入户访问有更多的限制，如时间上的限制，这时的问卷就不能询问较长的、复杂的问题。

再如，在电话访问中，被调查者可以与调查员交谈，但是看不到问卷，这就决定了只能问一些短的和比较简单的问题。邮寄问卷是自己独自填写的，被调查者与调研者没有直接的交流，因此问题也应简单些并要给出详细的指导语。

3) 确定每个问答题的内容

决定了访问方法的类型，下一步就是确定每个问答题的内容，即每个问答题应包括什么。

每个问答题的内容设计满足以下两项原则。

(1) 必要性。必要性是指问卷中每个问题要确定它的必要性，不要出现可有可无的问题。

(2) 目的性。目的性是指问卷中的每一个问答题都应对所需的信息有所贡献，或服务于某些特定的目的。如果从一个问答题得不到可以满意的使用数据，那么这个问答题就应该取消。

当然有些时候，还可以“故意”问一些与所需信息没有直接联系的问答题。这些没有直接联系的问答题通常放在问卷的开头，目的是能让被调查者乐于介入此调查中。

4) 决定问答题的结构

一般来说，调查问卷的问题有开放性问题和封闭性问题两种类型。

(1) 开放性问题是一种被调查者用他们自己的语言自由回答和解释有关想法的问题，调查者对问卷中的问题不具体提供选择答案的问题。例如，“您为什么喜欢××可乐的电视广告？”

开放性问题的优点是，提问比较简单，回答比较真实，即数据的质量较高。但它的缺点是，难以统计分析，即难以量化。因此，开放性问题在探索性调研中很有帮助，但在大规模的抽样调查中，它就弊大于利了。

(2) 封闭性问答题是指问卷调查中的问题要事先设计好各种可能的答案的问题，由被调查者从备选的答案中选定一个或几个即可。

例如，您购买住房时考虑的主要因素是什么？

A. 周边环境

B. 价格

C. 交通情况

D. 面积

E. 施工质量

F. 格局

由于答案标准化，因此封闭性问答题的优点是回答方便，有利于提高问卷的回收率；

易于进行各种统计处理和统计分析。其缺点是被调查者只能在规定的范围内回答，无法表达自己的真实想法，即存在着一定的偏差。所以此问答题的方式只适用于收集被调查者已经有明确看法的意向调查，不适用于初步探索性调查。

5) 决定问题的措辞

问题的措辞是指将想要的问题内容和结构，翻译成调查对象可以清楚而轻松地理解的用语。主要包括以下几点。

(1) 问题的措辞要求多用普通用语、语法，如果必须要用专业术语，必须对其加以解释。

(2) 要避免一句话中使用两个以上的同类概念或双重否定语。

(3) 要防止诱导性、暗示性的问题，以免影响回卷者的思考。

(4) 问及敏感性的问题时要讲究技巧。

(5) 行文要浅显易读，要考虑到回卷者的知识水准及文化程度，不要超过回卷者的领悟能力。

(6) 可运用方言，访问时更是如此。

6) 安排问题的顺序

通常问卷的问题安排顺序为先浅后深、先简后繁、先具体后抽象，所以最初安排的问题应容易回答且具有趣味性，旨在提高应答者的兴趣。核心问题往往置于问卷中间部分。即问卷中问题的顺序一般按下列规则排列。

(1) 容易回答的问题放前面，较难回答的问题放稍后，困窘性问题放后面，个人资料的事实性问题放卷首。

(2) 封闭性问题放前面，开放性问题放后面。因为开放性问题往往需要时间来考虑答案和组织语言，放在前面会引起应答者的厌烦情绪，一旦出现这种情况，应答者会中途放弃。

(3) 要注意问题的逻辑顺序，按时间顺序、类别顺序等合理排列。

7) 确定格式和排版

问题的格式及问卷的排版都会对结果产生显著的影响。格式有 3 种：行式排列、列式排列和矩阵式排列。

(1) 行式排列，即将所有备选项排成一行的排列方式。

例如，您购买住房时考虑的主要因素是什么？

A. 价格　　B. 面积　　C. 交通情况　　D. 周边环境　　E. 格局

(2) 列式排列，即将所有备选项排成一列，放在每个问题下边的排列方式。

例如，您购买住房时考虑的主要因素是什么？

A. 价格　　B. 面积　　C. 交通情况　　D. 周边环境

E. 格局　　F. 施工质量

(3) 矩阵式排列，即当多个问题具有相同的选项时，可将其设计成矩阵式。

例如，依您对下列问题的同意程度进行适当的选择：

	非常同意	同意	中立	不同意	非常不同意
① 粮食价格应降低 10%	[ ]	[ ]	[ ]	[ ]	[ ]
② 粮食价格应保持稳定	[ ]	[ ]	[ ]	[ ]	[ ]
③ 粮食价格应提高 10%以下	[ ]	[ ]	[ ]	[ ]	[ ]
④ 粮食价格应提高 10%～20%	[ ]	[ ]	[ ]	[ ]	[ ]

8) 拟定问卷的初稿和预调查

以上的程序全部完成，即完成了问卷的初稿。拟定好问卷的初稿后，得到管理层的认可后，必须进行预调查。

预调查要以最终调查的相同形式进行，如果调查是入户调查，预调查也应当采用入户的形式。在预调查完成后，任何需要改变的地方都应当切实修改。在进行实地调查前，问卷应当再一次获得各方的认可，如果预调查导致问卷产生较大的改动，需要进行第二次的预调查。

9) 设计正式问卷

问卷表的一般结构包括标题、说明、主体、致谢语 4 项。

(1) 标题。每份问卷都有一个研究主题。研究者要为此次调查定个题目，反映这个研究主题，使人一目了然，增强填答者的兴趣和责任感。

(2) 说明。说明可以让被调查者了解调查的目的和内容，也可以是指导语，说明这个调查的目的和意义，填答问卷的要求和注意事项，下面同时填上调查单位名称和年月日。说明通常放在问卷的前面，篇幅宜小不宜大，一般以两三百字为宜。

(3) 主体。主体是指问卷的核心部分，即问题和答案。从形式上看，问题可分为开放式和封闭式两种。从内容上看，可以分为事实性问题、意见性问题、困窘性问题等 (详见下面的 3 种问卷设计技巧)。

(4) 致谢语。为了表示对调查对象真诚合作的协议，研究者应当在问卷的末端写上感谢的话。但如果前面的说明中已包含感谢的话语，末端可不用。

在问卷设计完成后，进入问卷调查的实施及数据的回收和整理(此内容将在 2.4 节中详细讲解)。

3. 问卷设计的技巧

1) 事实性问题

事实性问题主要是要求应答者回答一些有关事实的问题。主要目的在于求取事实资料。

市场调查中，许多问题均属事实性问题，如被调查者的个人资料：性别、年龄、职业、收入、家庭状况、居住环境、教育程度等。这些问题又称为“分类性问题”，因为可根据所获得的资料而将应答者分类。在问卷设计之中，通常将事实性问题放在前面。

2) 意见性问题

在问卷中，往往会询问应答者一些有关意见或态度的问题。例如：“你是否喜欢××品牌饮料？”

意见性问题事实上即态度调查问题，关键在于被调查者是否愿意表达他真正的态度。这种问题通常有两种方法提问：一种方法是对意见性问题的答案只用百分比表示；另一种方法则旨在衡量应答者的态度，故可将答案化成分数。

3) 困窘性问题

困窘性问题是指应答者不愿在调查员面前作答的某些问题，如关于私人的问题，或不为一般社会道德所接纳的行为、态度的问题，但由于调查的需要必须获得困窘性问题的答案。为避免被调查者在应答时做不真实回答，可采用以下方法：

(1) 间接问题法。是指那些不宜于直接回答，而采用间接提问的方式得到所需答案的问题。这种提问法考虑到被调查者的顾虑。例如，“你同他们的看法是否一样？”

(2) 断定性问题。有些问题是先假定应答者已有该种态度或行为。例如，“你每天抽多

少支香烟？”事实上该应答者极可能根本不抽烟，这种问题则为断定性问题。正确处理这种问题的方法是在断定性问题之前加一问题，如“你抽烟吗？”，如果应答者回答“是”，用断定性问题继续问下去才有意义，否则在过滤问题后就应停止。

(3) 假设性问题。即通过假设某一情景或现象存在而向被调查者提出的问题。例如，“如果××矿泉水涨至 3 元，你是否将改喝未涨价的饮料？”

2.3 统计数据质量

统计数据的质量直接影响到统计分析的结论。为确保统计数据的质量，在数据收集、整理、分析各阶段都应尽可能减小误差，尤其是在数据收集阶段。

2.3.1 统计数据的误差

统计数据的误差是指统计数据与客观现实之间的差距。统计数据的误差主要包括抽样误差和非抽样误差。

抽样误差是指用样本推断总体时可能产生的误差。这种误差的产生原因一般有以下几种。

(1) 由于抽取样本时没有遵循随机原则而产生的。

(2) 由于样本结构与总体结构的差异而产生的。

(3) 由于样本容量不足而产生的。

抽样误差是无法消除的，但在调查方案设计前可以事先进行控制或计算。例如，按研究要求的精度，利用公式可以计算出最小样本容量。

非抽样误差是指在调查过程中由于调查者或被调查者的人为因素所造成的误差。例如，调查者在调查过程中的填报错误、抄录错误、汇总错误等引起误差，属于非抽样误差；再如，被调查者的故意虚报或瞒报引起的误差，也属于非抽样误差。对于非抽样误差来说，无论采用哪种方式调查都有可能产生。

2.3.2 统计数据的质量要求

一般统计数据的质量评价标准主要有 6 个方面。

(1) 精度：最低的抽样误差或随机误差。

(2) 准确性：最小的非抽样误差或偏差。

(3) 关联性：满足用户决策、管理和研究的需要。

(4) 及时性：在最短的时间里取得并公布数据。

(5) 一致性：保证时间序列的可比性。

(6) 最低成本：在满足以上标准的前提下，以最经济的方式取得数据。

2.3.3 降低统计数据误差的措施

1. 非抽样误差减小的措施

对于非抽样误差，必须采取各种措施，降低或减小可能发生的各种非抽样误差，把它缩小到最低限度范围内。主要的措施有以下两种。

(1) 正确制定好严密的调查方案，详细界定各种调查项目和计算方法。

(2) 切实落实好调查方案的各项内容。

2. 抽样误差减小的措施

要减小抽样误差，一般情况下，可在选择抽样调查单位环节下手。例如，在抽样调查时选取有代表性的调查单位；抽样调查遵循随机原则；在抽样之前确定好样本容量，并保证不随意更换样本单位等。

2.4 案例分析：啤酒市场的调查与分析及 Excel 上机应用——数据的收集

2.4.1 调查问卷的设计

假设此次调查采用电子邮件的方式进行，调查目前吉林省的几个城市消费者饮用啤酒的行为习惯。从经济的角度我们知道与消费者的行为有关的包括消费者的性别、年龄、学历和居住城市等个人信息，除此之外，还包括消费者对啤酒的一些看法等。根据这些内容我们首先设计调查问卷。

1. 设计问卷的说明和标题

根据前面的介绍，说明不是问卷的主要部分，这部分主要是对调查目的、意义及填写要求的说明。标题反映这个研究主题，使人一目了然，增强填答者的兴趣和责任感。根据研究目的，设计了啤酒消费者行为调查表，如图 2.1 所示。

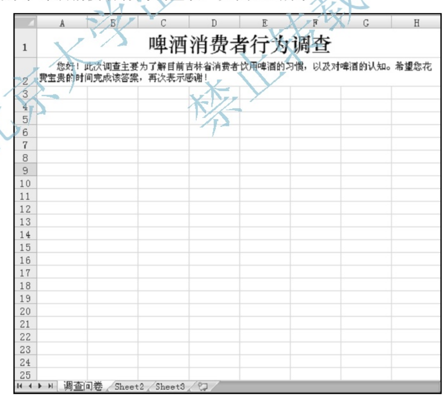


图 2.1 啤酒消费者行为调查表

2. 设计问卷的主体

问卷主体是市场调查所要收集的主要信息，它由一个个问题及相应的选择项目组成，通过主体部分问题的设计和被调查者的答复，市场调查者可以对被调查者的基本个人情况和对某一特定事物的态度、意见倾向及行为有较充分的了解。设计步骤如下(本书案例的操作全部使用最普及的 Excel 2007 进行讲解)。

第一步：在“调查问卷”工作表中输入如图 2.2 所示的问卷题目。

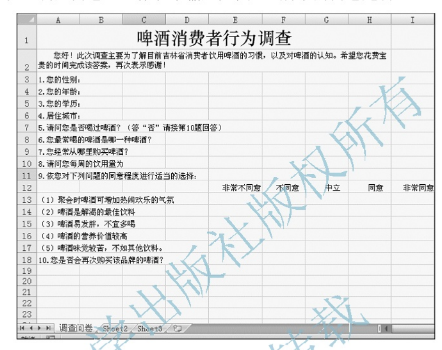


图 2.2　问卷题目示意图

第二步：为了美观，取消网线格，操作过程，单击“视图”选项卡，取消勾选“网格线”复选框，即取消网格线，得到的结果如图 2.3 所示。

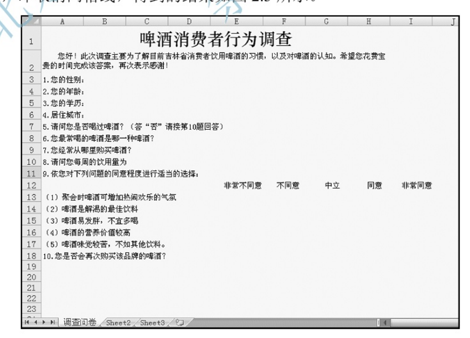
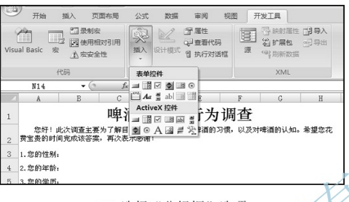
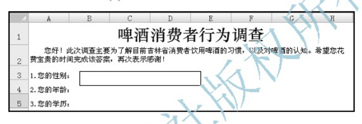


图 2.3　取消网格后的问卷题目示意图

第三步：单击“开发工具”→“插入”的下拉按钮，在弹出的下拉列表中选择“分组框”选项，然后拖动鼠标，在工作表中适当的位置创建分组框，如图 2.4 所示。



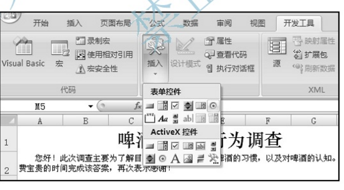
(a) 选择“分组框”选项



(b) 分组框效果

图 2.4　创建分组框

第四步：单击“开发工具”→“插入”的下拉按钮，在弹出的下拉列表中选择“选项按钮”选项，在上一步中绘制的分组框内部绘制选项按钮，并将选项按钮的名称更改为“男”；之后按住 Ctrl 键，复制一个选项按钮，仍然将该选项按钮放置于分组框内，并将名称更改为“女”，如图 2.5 所示。



(a) 选择“选项按钮”选项

啤酒消费者行为调查

您好！此次调查主要为了解目前吉林省消费者饮用啤酒的习惯，以及对啤酒的认知
费宝贵的时间完成该答案，再次表示感谢！

1.您的性别：　○男　○女

2.您的年龄：

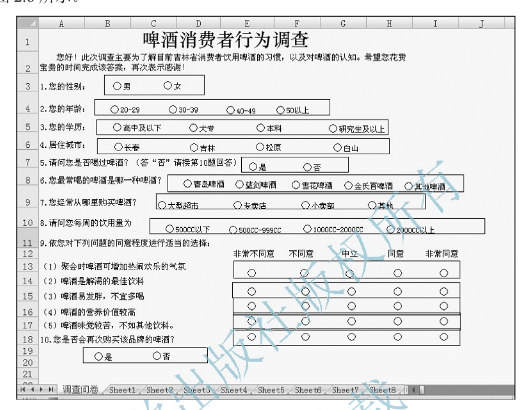
3.您的学历：

(b) 选项按钮效果

图 2.5　绘制选项按钮

此时当指向选项按钮时，鼠标指针的形状会显示为手状，单击，将选中当前选项按钮。

第五步：按照上面的顺序，依次为问卷其他问题设计好选项，最后得出的问卷题目及选项如图 2.6 所示。



啤酒消费者行为调查

您好！此次调查主要为了解目前吉林省消费者饮用啤酒的习惯，以及对啤酒的认知。希望您花费宝贵的时间完成该答案，再次表示感谢！

1.您的性别： ○男 ○女

2.您的年龄： ○20-29 ○30-39 ○40-49 ○50以上

3.您的学历： ○高中及以下 ○大专 ○本科 ○研究生及以上

4.居住城市： ○长春 ○吉林 ○松原 ○白山

5.请问您是否喝过啤酒？（答“否”请接第10题回答） ○是 ○否

6.您最常喝的啤酒是哪一种啤酒？ ○青岛啤酒 ○蓝剑啤酒 ○雪花啤酒 ○金氏百啤酒 ○其他啤酒

7.您经常从哪里购买啤酒？ ○大型超市 ○专卖店 ○小卖部 ○其他

8.请问您每周的饮用量为 ○500CC以下 ○500CC-999CC ○1000CC-2000CC ○2000CC以上

9.依您对下列问题的同意程度进行适当的选择：

	非常不同意	不同意	中立	同意	非常同意
（1）聚会时啤酒可增加热闹欢乐的气氛	○	○	○	○	○
（2）啤酒是解渴的最佳饮料	○	○	○	○	○
（3）啤酒易发胖，不宜多喝	○	○	○	○	○
（4）啤酒的营养价值较高	○	○	○	○	○
（5）啤酒味觉较苦，不如其他饮料。	○	○	○	○	○

10.您是否会再次购买该品牌的啤酒？ ○是 ○否

调查问卷 Sheet1 Sheet2 Sheet3 Sheet4 Sheet5 Sheet6 Sheet7 Sheet8

图 2.6 新设计的调查问卷示意图

第六步：设计“提交问卷”按钮，单击“插入”→“形状”的下拉按钮，在弹出的下拉列表中选择“矩形”选项，拖动鼠标，在问卷的空白处绘制矩形，然后右击矩形，在弹出的快捷菜单中选择“添加文字”选项，添加“提交问卷”文字，如图 2.7 所示。

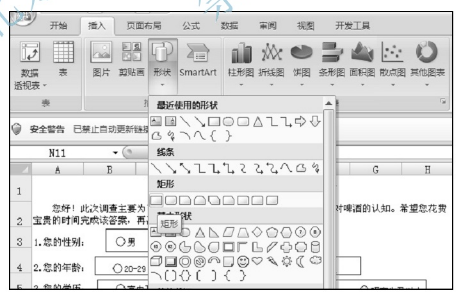


图 2.7 设计“提交”问卷按钮

最后，得到最终问卷，如图 2.8 所示(这里主要介绍方法，所以问卷就不加以设计背景了)。

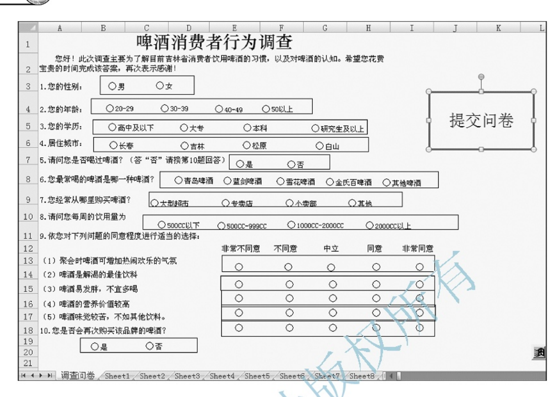


图 2.8 最终问卷示意图

2.4.2 自动接收问卷结果的设置

此次调查是以电子邮件的方式进行的，为了避免手动输入调查结果，可以通过设置自动接收问卷结果来实现自动化。本书将利用 VAB 代码自动记录所有调查结果的方法。

具体的创建方法如下。

第一步：建立一个新的工作表，命名为“自动统计调查结果”，并在其上创建如图 2.9 所示的问卷题目信息。

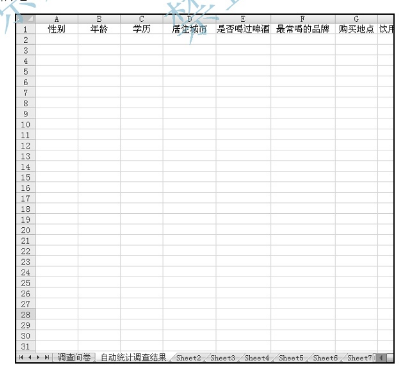


图 2.9 “自动统计调查结果”工作表

第二步：切换到“调查问卷”工作表，右击单选按钮“男”，在弹出的快捷菜单中选择“设置控件格式”选项，如图 2.10 所示，弹出“设置控件格式”对话框。

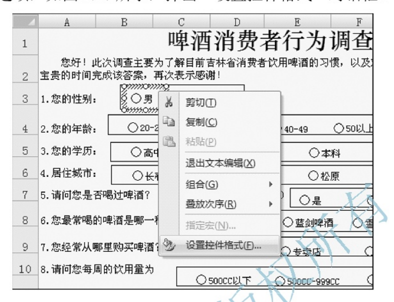


图 2.10 “设置控件格式”选项

第三步：在“设置控件格式”对话框中，单击“控制”选项卡，然后单击“单元格链接”文本框右侧的单元格引用按钮，如图 2.11 所示。



图 2.11 “设置控件格式”对话框

第四步：此时，“设置控件格式”对话框会自动折叠为只显示一个文本框，单击“自动统计调查结果”工作表，然后选定“自动统计调查结果”工作表中的 A2 单元格，此时该单元格的引用路径会显示在对话框的文本框中，如图 2.12 所示。



图 2.12 单元格的引用路径显示结果

第五步：在选定引用的单元格后，单击上一步中对话框的“关闭”按钮，返回“设置对象格式”对话框，然后关闭“设置对象格式”对话框。

此时，在“调查问卷”工作表选中选项按钮“男”后，切换到“自动统计调查结果”工作表，会发现该工作表的 A2 单元格中自动出现数值“1”；如果在“调查问卷”工作表选中选项按钮“女”后，切换到“自动统计调查结果”工作表，会发现该工作表的 A2 单元格中自动出现数值“2”如图 2.13 所示。



图 2.13　自动接收问卷结果

第六步：用同样的方法为其余的题目设置备选择答案。并将各组题目的选项按钮和对应的“自动统计调查结果”工作表中的单元格链接。

2.4.3　“自动统计调查结果”工作表的隐藏和问卷邮件的发送

制作好调查问卷后，接下来要做的工作就是将问卷发送到被调查者的邮箱。但是，由于被调查者只需要看到“调查问卷”工作表，所以在发送邮件之前，还需要进行隐藏“自动统计调查结果”工作表并保护工作簿。具体操作如下。

第一步：切换到“自动统计调查结果”工作表，右击该工作表，在弹出的快捷菜单中选择“隐藏”选项，如图 2.14 所示。



图 2.14　选择“隐藏”选项

此时，“自动统计调查结果”工作表就自动隐藏了，如图 2.15 所示。

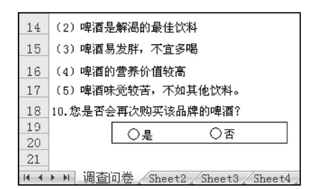


图 2.15 “自动统计调查结果”工作表自动隐藏

第二步：切换到“调查问卷”工作表，对该工作表进行保护。单击“审阅”→“保护工作簿”的下拉按钮，在弹出的下拉列表中选择“保护结构和窗口”选项，弹出“保护结构和窗口”对话框，在“密码”文本框中输入密码，如“123”，如图 2.16 所示。

第三步：单击“确定”按钮，弹出“确认密码”对话框，再次输入密码，如图 2.17 所示，然后单击“确定“按钮。



图 2.16 “保护结构和窗口”对话框

图 2.17 “确认密码”对话框

此时，就可以把调查问卷发送给被调查者。

2.4.4 调查结果资料库的创建

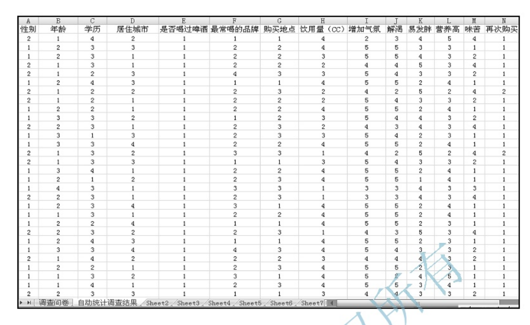
设此次调查共发送 50 份电子邮件，收到问卷结果 35 份，其中 5 份问卷答题不符合要求，不予统计，下面把另外 30 份问卷结果进行统计。操作如下。

第一步：打开“调查问卷”工作表，单击“审阅”→“保护工作簿”的下拉按钮，在弹出的下拉列表中选择“保护结构和窗口”选项，弹出“撤销工作簿保护”对话框，在“密码”文本框中输入之前设置的密码，然后单击“确定”按钮，如图 2.18 所示。



图 2.18 “撤销工作簿保护”对话框

第二步：此时显示“自动统计调查结果”工作表，有效的 30 份问卷调查结果出现在该工作表中，如图 2.19 所示。

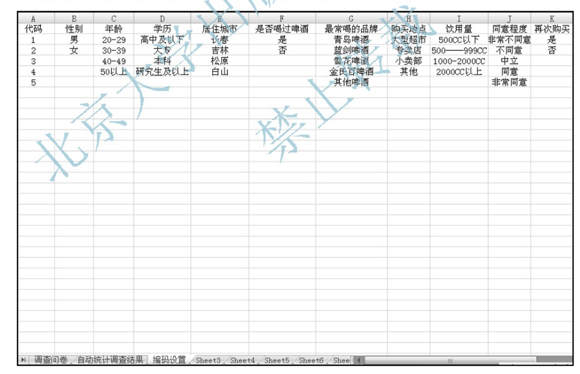


性别	年龄	学历	居住城市	是否喝过啤酒	最常喝的品牌	购买地点	饮用量（CC）	增加气氛	解渴	易发胖	营养高	味苦	再次购买
2	1	4	2	1	1	1	4	2	3	4	5	4	1
1	2	3	3	1	2	2	4	5	5	3	3	1	1
1	2	3	1	1	2	2	3	5	5	4	3	2	1
2	1	3	1	1	2	2	2	4	4	5	3	4	1
2	1	2	3	1	4	3	3	5	4	3	3	2	1
1	2	4	3	1	1	1	4	5	5	2	4	1	1
2	1	2	2	1	2	3	2	4	2	5	2	4	2
2	1	2	1	1	2	2	2	5	4	3	3	2	1
1	2	2	1	1	2	2	4	5	5	2	4	1	1
1	3	3	2	1	1	2	3	5	4	4	3	2	1
2	2	3	1	1	2	3	2	4	3	4	3	4	1
1	3	1	3	1	2	3	3	5	4	2	3	1	1
1	3	3	4	1	2	2	4	5	5	2	4	1	1
2	1	3	2	1	3	3	1	4	2	5	2	4	2
2	1	3	3	1	1	1	3	5	4	3	3	2	1
1	3	4	1	1	2	2	4	5	5	2	4	1	1
1	2	1	2	1	2	3	4	5	5	1	4	1	1
1	4	3	1	1	3	3	1	3	3	4	3	3	1
2	2	3	1	1	2	3	1	3	3	4	3	4	1
1	2	3	4	1	3	1	4	5	5	2	4	1	1
1	1	3	1	1	2	2	4	5	5	2	4	1	1
1	2	2	4	1	1	1	4	5	5	2	3	1	1
2	2	3	2	1	2	3	1	4	3	5	3	4	1
1	2	4	3	1	1	1	4	5	5	2	3	1	1
1	3	3	4	1	4	3	4	5	4	3	3	2	1
2	1	4	2	1	2	2	3	4	4	4	3	2	1
1	2	2	1	1	2	3	4	5	5	2	3	1	1
1	1	3	2	1	3	1	4	5	2	4	5	1	1
1	1	4	1	1	2	3	4	5	5	3	4	1	1
2	2	3	3	1	1	1	3	4	4	3	3	2	1

图 2.19　问卷调查结果

由图 2.19 可知，即每位受访者的答题结果均以数字代码的形式显示在“自动统计调查结果”工作表中，要把代码转换为具体的内容。

第三步：在“自动统计调查结果”工作表后插入一个新的工作表，命名为“编码设置”。在“编码设置”工作表中设置不同答案的编码，如图 2.20 所示。



代码	性别	年龄	学历	居住城市	是否喝过啤酒	最常喝的品牌	购买地点	饮用量	同意程度	再次购买
1	男	20-29	高中及以下	长春	是	青岛啤酒	大型超市	500CC以下	非常不同意	是
2	女	30-39	大专	吉林	否	蓝剑啤酒	专卖店	500——999CC	不同意	否
3		40-49	本科	松原		雪花啤酒	小卖部	1000-2000CC	中立	
4		50以上	研究生及以上	白山		金氏百啤酒	其他	2000CC以上	同意	
5						其他啤酒			非常同意	

图 2.20　编码设置

第四步：在“自动统计调查结果”工作表中，分别在每列数据右侧插入一个空白列，如图 2.21 所示。

	A	B	C	D	E	F	G
1	性别	性别	年龄	年龄	学历	学历	居住城
2	2		1		4		2
3	1		2		3		3
4	1		2		3		1
5	2		1		3		1
6	2		1		2		3
7	1		2		4		3
8	2		1		2		2
9	2		1		2		1
10	1		2		2		1
11	1		3		3		2
12	2		2		3		1
13	1		3		1		3
14	1		3		3		4
15	2		1		3		2
16	2		1		3		3
17	1		3		4		1
18	1		2		1		2

图 2.21　插入空白列

第五步：切换到“编码设置”工作表，选定单元格区域 A1:K6，然后单击“公式”→“定义名称”的下拉按钮，在弹出的下拉列表中选择“定义名称”选项，如图 2.22 所示。

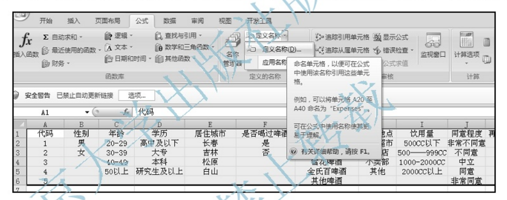


图 2.22　“编码设置”工作表操作示意图

第六步：在“定义名称”对话框中的“名称”文本框中输入名称，如 data，然后单击“确定”按钮，如图 2.23 所示。



图 2.23　输入名称

第七步：切换到“自动统计调查结果”工作表，在 B2 单元格中输入替代函数，即：“=VLOOKUP(A2,data,2,FALSE)”，此时 B2 单元格自动替换成“女”，如图 2.24 所示。

B2 f_x =VLOOKUP(A2, data, 2, FALSE)

	A	B	C	D	E	F
1	性别	性别	年龄	年龄	学历	学历
2	2	女	1		4	
3	1		2		3	
4	1		2		3	
5	2		1		3	
6	2		1		2	
7	1		2		4	

图 2.24　输入替代函数

第八步：拖动 B2 单元格右下角的填充柄向下复制公式，替换出每位受访者的性别信息，其结果如图 2.25 所示。

B2 f_x =VLOOKUP(A2, data, 2, FALSE)

	A	B	C	D	E	F
1	性别	性别	年龄	年龄	学历	学历
2	2	女	1		4	
3	1	男	2		3	
4	1	男	2		3	
5	2	女	1		3	
6	2	女	1		2	
7	1	男	2		4	
8	2	女	1		2	
9	2	女	1		2	
10	1	男	2		2	
11	1	男	3		3	
12	2	女	2		3	
13	1	男	3		1	
14	1	男	3		3	
15	2	女	1		3	
16	2	女	1		3	
17	1	男	3		4	
18	1	男	2		1	
19	1	男	4		3	
20	2	女	2		3	
21	1	男	2		3	
22	1	男	1		3	
23	1	男	2		2	
24	2	女	2		3	
25	1	男	2		4	
26	1	男	3		3	
27	2	女	1		4	
28	1	男	2		2	
29	1	男	1		3	
30	1	男	1		4	
31	2	女	2		3	

图 2.25　替换每位受访者的性别信息

第九步：在 D2 单元格中输入公式“=VLOOKUP(C2,data,3,FALSE)”，按 Enter 键后向下复制公式，其结果如图 2.26 所示。

第十步：在 F2 单元格中输入公式“=VLOOKUP(E2,data,4,FALSE)”，按 Enter 键后向下复制公式，其结果如图 2.27 所示。

=VLOOKUP(C2,data,3,FAL

性别	年龄	年龄	学历
女	1	20-29	4
男	2	30-39	3
男	2	30-39	3
女	1	20-29	3
女	1	20-29	2
男	2	30-39	4
女	1	20-29	2
女	1	20-29	2
男	2	30-39	2
男	3	40-49	3
女	2	30-39	3
男	3	40-49	1
男	3	40-49	3
女	1	20-29	3
女	1	20-29	3
男	3	40-49	4
男	2	30-39	1
男	4	50以上	3
女	2	30-39	3
男	2	30-39	3
男	1	20-29	3
男	2	30-39	2
女	2	30-39	3
男	2	30-39	4
男	3	40-49	3
女	1	20-29	4

问卷 自动统计调查结果 编码设置

图 2.26 输入公式“=VLOOKUP(C2,data,3,FALSE)”后的结果

=VLOOKUP(E2,data,4,FALSE)

年龄	年龄	学历	学历
1	20-29	4	研究生及以上
2	30-39	3	本科
2	30-39	3	本科
1	20-29	3	本科
1	20-29	2	大专
2	30-39	4	研究生及以上
1	20-29	2	大专
1	20-29	2	大专
2	30-39	2	大专
3	40-49	3	本科
2	30-39	3	本科
3	40-49	1	高中及以下
3	40-49	3	本科
1	20-29	3	本科
1	20-29	3	本科
3	40-49	4	研究生及以上
2	30-39	1	高中及以下
4	50以上	3	本科
2	30-39	3	本科
2	30-39	3	本科
1	20-29	3	本科
2	30-39	2	大专
2	30-39	3	本科
2	30-39	4	研究生及以上
3	40-49	3	本科
1	20-29	4	研究生及以上

统计调查结果 编码设置

图 2.27 输入公式“=VLOOKUP(E2,data,4,FALSE)”后的结果

第十一步：在 H2 单元格中输入公式“=VLOOKUP(G2,data,5,FALSE)”，按 Enter 键后向下复制公式，其结果如图 2.28 所示。

=VLOOKUP(G2,data,5,FALSE)

年龄	学历	学历	居住城市	居住城市
20-29	4	研究生及以上	2	吉林
30-39	3	本科	3	松原
30-39	3	本科	1	长春
20-29	3	本科	1	长春
20-29	2	大专	3	松原
30-39	4	研究生及以上	3	松原
20-29	2	大专	2	吉林
20-29	2	大专	1	长春
30-39	2	大专	1	长春
40-49	3	本科	2	吉林
30-39	3	本科	1	长春
40-49	1	高中及以下	3	松原
40-49	3	本科	4	白山
20-29	3	本科	2	吉林
20-29	3	本科	3	松原
40-49	4	研究生及以上	1	长春
30-39	1	高中及以下	2	吉林
50以上	3	本科	1	长春
30-39	3	本科	1	长春
30-39	3	本科	4	白山
20-29	3	本科	1	长春
30-39	2	大专	4	白山
30-39	3	本科	2	吉林
30-39	4	研究生及以上	3	松原
40-49	3	本科	4	白山
20-29	4	研究生及以上	2	吉林

问卷 自动统计调查结果 编码设置

图 2.28 输入公式“=VLOOKUP(G2,data,5,FALSE)”后的结果

第十二步：在 J2 单元格中输入公式“=VLOOKUP(I2,data,6,FALSE)”，按 Enter 键后向下复制公式，其结果如图 2.29 所示。

=VLOOKUP(I2, data, 6, FALSE)

D	E	F	G	H	I	J	K
年龄	学历	学历	居住城市	居住城市	是否喝过啤酒	是否喝过啤酒	最常喝的品牌
20-29	4	研究生及以上	2	吉林	1	是	1
30-39	3	本科	3	松原	1	是	2
30-39	3	本科	1	长春	1	是	2
20-29	3	本科	1	长春	1	是	2
20-29	2	大专	3	松原	1	是	4
30-39	4	研究生及以上	3	松原	1	是	1
20-29	2	大专	2	吉林	1	是	2
20-29	2	大专	1	长春	1	是	2
30-39	2	大专	1	长春	1	是	2
40-49	3	本科	2	吉林	1	是	1
30-39	3	本科	1	长春	1	是	2
40-49	1	高中及以下	3	松原	1	是	2
40-49	3	本科	4	白山	1	是	2
20-29	3	本科	2	吉林	1	是	3
20-29	3	本科	3	松原	1	是	1
40-49	4	研究生及以上	1	长春	1	是	2
30-39	1	高中及以下	2	吉林	1	是	2
50以上	3	本科	1	长春	1	是	3
30-39	3	本科	1	长春	1	是	2
30-39	3	本科	4	白山	1	是	3
20-29	3	本科	1	长春	1	是	2
30-39	2	大专	4	白山	1	是	1
30-39	3	本科	2	吉林	1	是	2
30-39	4	研究生及以上	3	松原	1	是	1
40-49	3	本科	4	白山	1	是	4
20-29	4	研究生及以上	2	吉林	1	是	2

问卷 自动统计调查结果 编码设置

图 2.29　输入公式“=VLOOKUP(I2,data,6,FALSE)”后的结果

第十三步：在 L2 单元格中输入公式“=VLOOKUP(K2,data,7,FALSE)”，按 Enter 键后向下复制公式，其结果如图 2.30 所示。

=VLOOKUP(K2, data, 7, FALSE)

D	E	F	G	H	I	J	K	L
年龄	学历	学历	居住城市	居住城市	是否喝过啤酒	是否喝过啤酒	最常喝的品牌	最常喝的品牌
20-29	4	研究生及以上	2	吉林	1	是	1	青岛啤酒
30-39	3	本科	3	松原	1	是	2	蓝剑啤酒
30-39	3	本科	1	长春	1	是	2	蓝剑啤酒
20-29	3	本科	1	长春	1	是	2	蓝剑啤酒
20-29	2	大专	3	松原	1	是	4	金氏百啤酒
30-39	4	研究生及以上	3	松原	1	是	1	青岛啤酒
20-29	2	大专	2	吉林	1	是	2	蓝剑啤酒
20-29	2	大专	1	长春	1	是	2	蓝剑啤酒
30-39	2	大专	1	长春	1	是	2	蓝剑啤酒
40-49	3	本科	2	吉林	1	是	1	青岛啤酒
30-39	3	本科	1	长春	1	是	2	蓝剑啤酒
40-49	1	高中及以下	3	松原	1	是	2	蓝剑啤酒
40-49	3	本科	4	白山	1	是	2	蓝剑啤酒
20-29	3	本科	2	吉林	1	是	3	雪花啤酒
20-29	3	本科	3	松原	1	是	1	青岛啤酒
40-49	4	研究生及以上	1	长春	1	是	2	蓝剑啤酒
30-39	1	高中及以下	2	吉林	1	是	2	蓝剑啤酒
50以上	3	本科	1	长春	1	是	3	雪花啤酒
30-39	3	本科	1	长春	1	是	2	蓝剑啤酒
30-39	3	本科	4	白山	1	是	3	雪花啤酒
20-29	3	本科	1	长春	1	是	2	蓝剑啤酒
30-39	2	大专	4	白山	1	是	1	青岛啤酒
30-39	3	本科	2	吉林	1	是	2	蓝剑啤酒
30-39	4	研究生及以上	3	松原	1	是	1	青岛啤酒
40-49	3	本科	4	白山	1	是	4	金氏百啤酒
20-29	4	研究生及以上	2	吉林	1	是	2	蓝剑啤酒

问卷 自动统计调查结果 编码设置

图 2.30　输入公式“=VLOOKUP(K2,data,7,FALSE)”后的结果

第十四步：在 N2 单元格中输入公式“=VLOOKUP(M2,data,8,FALSE)”，按 Enter 键后向下复制公式，其结果如图 2.31 所示。

=VLOOKUP(M2,data,8,FALSE)

J	K	L	M	N	O
是否喝过啤酒	最常喝的品牌	最常喝的品牌	购买地点	购买地点	饮用量（CC）
是	1	青岛啤酒	1	大型超市	4
是	2	蓝剑啤酒	2	专卖店	4
是	2	蓝剑啤酒	2	专卖店	3
是	2	蓝剑啤酒	2	专卖店	2
是	4	金氏百啤酒	3	小卖部	3
是	1	青岛啤酒	1	大型超市	4
是	2	蓝剑啤酒	3	小卖部	2
是	2	蓝剑啤酒	2	专卖店	2
是	2	蓝剑啤酒	2	专卖店	4
是	1	青岛啤酒	2	专卖店	3
是	2	蓝剑啤酒	3	小卖部	2
是	2	蓝剑啤酒	3	小卖部	3
是	2	蓝剑啤酒	2	专卖店	4
是	3	雪花啤酒	3	小卖部	1
是	1	青岛啤酒	1	大型超市	3
是	2	蓝剑啤酒	2	专卖店	4
是	2	蓝剑啤酒	3	小卖部	4
是	3	雪花啤酒	3	小卖部	1
是	2	蓝剑啤酒	3	小卖部	1
是	3	雪花啤酒	1	大型超市	4
是	2	蓝剑啤酒	2	专卖店	4
是	1	青岛啤酒	1	大型超市	4
是	2	蓝剑啤酒	3	小卖部	1
是	1	青岛啤酒	1	大型超市	4
是	4	金氏百啤酒	3	小卖部	4
是	2	蓝剑啤酒	2	专卖店	3

图 2.31　输入公式“=VLOOKUP(M2,data,8,FALSE)”后的结果

第十五步：在 P2 单元格中输入公式“=VLOOKUP(O2,data,9,FALSE)”，按 Enter 键后向下复制公式，其结果如图 2.32 所示。

=VLOOKUP(O2,data,9,FALSE)

I	J	K	L	M	N	O	P
是否喝过啤酒	是否喝过啤酒	最常喝的品牌	最常喝的品牌	购买地点	购买地点	饮用量（CC）	饮用量（CC）
1	是	1	青岛啤酒	1	大型超市	4	2000CC以上
1	是	2	蓝剑啤酒	2	专卖店	4	2000CC以上
1	是	2	蓝剑啤酒	2	专卖店	3	1000-2000CC
1	是	2	蓝剑啤酒	2	专卖店	2	500——999CC
1	是	4	金氏百啤酒	3	小卖部	3	1000-2000CC
1	是	1	青岛啤酒	1	大型超市	4	2000CC以上
1	是	2	蓝剑啤酒	3	小卖部	2	500——999CC
1	是	2	蓝剑啤酒	2	专卖店	2	500——999CC
1	是	2	蓝剑啤酒	2	专卖店	4	2000CC以上
1	是	1	青岛啤酒	2	专卖店	3	1000-2000CC
1	是	2	蓝剑啤酒	3	小卖部	2	500——999CC
1	是	2	蓝剑啤酒	3	小卖部	3	1000-2000CC
1	是	2	蓝剑啤酒	2	专卖店	4	2000CC以上
1	是	3	雪花啤酒	3	小卖部	1	500CC以下
1	是	1	青岛啤酒	1	大型超市	3	1000-2000CC
1	是	2	蓝剑啤酒	2	专卖店	4	2000CC以上
1	是	2	蓝剑啤酒	3	小卖部	4	2000CC以上
1	是	3	雪花啤酒	3	小卖部	1	500CC以下
1	是	2	蓝剑啤酒	3	小卖部	1	500CC以下
1	是	3	雪花啤酒	1	大型超市	4	2000CC以上
1	是	2	蓝剑啤酒	2	专卖店	4	2000CC以上
1	是	1	青岛啤酒	1	大型超市	4	2000CC以上
1	是	2	蓝剑啤酒	3	小卖部	1	500CC以下
1	是	1	青岛啤酒	1	大型超市	4	2000CC以上
1	是	4	金氏百啤酒	3	小卖部	4	2000CC以上
1	是	2	蓝剑啤酒	2	专卖店	3	1000-2000CC

图 2.32　输入公式“=VLOOKUP(O2,data,9,FALSE)”后的结果

第十六步：在 R2 单元格中输入公式“=VLOOKUP(Q2,data,10,FALSE)”，按 Enter 键后向下复制公式，其结果如图 2.33 所示。

=VLOOKUP(Q2, data, 10, FALSE)

N	O	P	Q	R	S
实地点	饮用量（CC）	饮用量（CC）	增加气氛	增加气氛	解渴
型超市	4	2000CC以上	2	不同意	3
卖店	4	2000CC以上	5	非常同意	5
卖店	3	1000-2000CC	5	非常同意	5
卖店	2	500——999CC	4	同意	4
卖部	3	1000-2000CC	5	非常同意	4
型超市	4	2000CC以上	5	非常同意	5
卖部	2	500——999CC	4	同意	2
卖店	2	500——999CC	5	非常同意	4
卖店	4	2000CC以上	5	非常同意	5
卖店	3	1000-2000CC	5	非常同意	4
卖部	2	500——999CC	4	同意	3
卖部	3	1000-2000CC	5	非常同意	4
卖店	4	2000CC以上	5	非常同意	5
卖部	1	500CC以下	4	同意	2
型超市	3	1000-2000CC	5	非常同意	4
卖店	4	2000CC以上	5	非常同意	5
卖部	4	2000CC以上	5	非常同意	5
卖部	1	500CC以下	3	中立	3
卖部	1	500CC以下	3	中立	3
型超市	4	2000CC以上	5	非常同意	5
卖店	4	2000CC以上	5	非常同意	5
型超市	4	2000CC以上	5	非常同意	5
卖部	1	500CC以下	4	同意	3
型超市	4	2000CC以上	5	非常同意	5
卖部	4	2000CC以上	5	非常同意	4
卖店	3	1000-2000CC	4	同意	4

自动统计调查结果 编码设置

图 2.33 输入公式“=VLOOKUP(Q2,data,10,FALSE)”后的结果

第十七步：在 T2 单元格中输入公式“=VLOOKUP(S2,data,10,FALSE)”，按 Enter 键后向下复制公式，其结果如图 2.34 所示。



=VLOOKUP(S2, data, 10, FALSE)

	O	P	Q	R	S	T
点	饮用量（CC）	饮用量（CC）	增加气氛	增加气氛	解渴	解渴
市	4	2000CC以上	2	不同意	3	中立
店	4	2000CC以上	5	非常同意	5	非常同意
店	3	1000-2000CC	5	非常同意	5	非常同意
店	2	500——999CC	4	同意	4	同意
部	3	1000-2000CC	5	非常同意	4	同意
市	4	2000CC以上	5	非常同意	5	非常同意
部	2	500——999CC	4	同意	2	不同意
店	2	500——999CC	5	非常同意	4	同意
店	4	2000CC以上	5	非常同意	5	非常同意
店	3	1000-2000CC	5	非常同意	4	同意
部	2	500——999CC	4	同意	3	中立
部	3	1000-2000CC	5	非常同意	4	同意
店	4	2000CC以上	5	非常同意	5	非常同意
部	1	500CC以下	4	同意	2	不同意
市	3	1000-2000CC	5	非常同意	4	同意
店	4	2000CC以上	5	非常同意	5	非常同意
部	4	2000CC以上	5	非常同意	5	非常同意
部	1	500CC以下	3	中立	3	中立
部	1	500CC以下	3	中立	3	中立
市	4	2000CC以上	5	非常同意	5	非常同意
店	4	2000CC以上	5	非常同意	5	非常同意
市	4	2000CC以上	5	非常同意	5	非常同意
部	1	500CC以下	4	同意	3	中立
市	4	2000CC以上	5	非常同意	5	非常同意
部	4	2000CC以上	5	非常同意	4	同意
店	3	1000-2000CC	4	同意	4	同意

动统计调查结果 编码设置

图 2.34 输入公式“=VLOOKUP(S2,data,10,FALSE)”后的结果

第十八步：在 V2 单元格中输入公式“=VLOOKUP(U2,data,10,FALSE)”，按 Enter 键后向下复制公式，其结果如图 2.35 所示。

=VLOOKUP(U2, data, 10, FALSE)

Q	R	S	T	U	V	W
增加气氛	增加气氛	解渴	解渴	易发胖	易发胖	营养
2	不同意	3	中立	4	同意	5
5	非常同意	5	非常同意	3	中立	3
5	非常同意	5	非常同意	4	同意	3
4	同意	4	同意	5	非常同意	3
5	非常同意	4	同意	3	中立	3
5	非常同意	5	非常同意	2	不同意	4
4	同意	2	不同意	5	非常同意	2
5	非常同意	4	同意	3	中立	3
5	非常同意	5	非常同意	2	不同意	4
5	非常同意	4	同意	4	同意	3
4	同意	3	中立	4	同意	3
5	非常同意	4	同意	2	不同意	3
5	非常同意	5	非常同意	2	不同意	4
4	同意	2	不同意	5	非常同意	2
5	非常同意	4	同意	3	中立	3
5	非常同意	5	非常同意	2	不同意	4
5	非常同意	5	非常同意	1	非常不同意	4
3	中立	3	中立	4	同意	3
3	中立	3	中立	4	同意	3
5	非常同意	5	非常同意	2	不同意	4
5	非常同意	5	非常同意	2	不同意	4
5	非常同意	5	非常同意	2	不同意	3
4	同意	3	中立	5	非常同意	3
5	非常同意	5	非常同意	2	不同意	3
5	非常同意	4	同意	3	中立	3
4	同意	4	同意	4	同意	3

自动统计调查结果　编码设置

图 2.35　输入公式“=VLOOKUP(U2,data,10,FALSE)”后的结果

第十九步：在 X2 单元格中输入公式“=VLOOKUP(W2,data,10,FALSE)”，按 Enter 键后向下复制公式，其结果如图 2.36 所示。

=VLOOKUP(W2, data, 10, FALSE)

Q	R	S	T	U	V	W	X
增加气氛	增加气氛	解渴	解渴	易发胖	易发胖	营养高	营养高
2	不同意	3	中立	4	同意	5	非常同意
5	非常同意	5	非常同意	3	中立	3	中立
5	非常同意	5	非常同意	4	同意	3	中立
4	同意	4	同意	5	非常同意	3	中立
5	非常同意	4	同意	3	中立	3	中立
5	非常同意	5	非常同意	2	不同意	4	同意
4	同意	2	不同意	5	非常同意	2	不同意
5	非常同意	4	同意	3	中立	3	中立
5	非常同意	5	非常同意	2	不同意	4	同意
5	非常同意	4	同意	4	同意	3	中立
4	同意	3	中立	4	同意	3	中立
5	非常同意	4	同意	2	不同意	3	中立
5	非常同意	5	非常同意	2	不同意	4	同意
4	同意	2	不同意	5	非常同意	2	不同意
5	非常同意	4	同意	3	中立	3	中立
5	非常同意	5	非常同意	2	不同意	4	同意
5	非常同意	5	非常同意	1	非常不同意	4	同意
3	中立	3	中立	4	同意	3	中立
3	中立	3	中立	4	同意	3	中立
5	非常同意	5	非常同意	2	不同意	4	同意
5	非常同意	5	非常同意	2	不同意	4	同意
5	非常同意	5	非常同意	2	不同意	3	中立
4	同意	3	中立	5	非常同意	3	中立
5	非常同意	5	非常同意	2	不同意	3	中立
5	非常同意	4	同意	3	中立	3	中立
4	同意	4	同意	4	同意	3	中立

问卷　自动统计调查结果　编码设置

图 2.36　输入公式“=VLOOKUP(W2,data,10,FALSE)”后的结果

第二十步：在 Z2 单元格中输入公式“=VLOOKUP(Y2,data,10,FALSE)”，按 Enter 键后向下复制公式，其结果如图 2.37 所示。

=VLOOKUP(Y2,data,10,FALSE)

U	V	W	X	Y	Z	AA
易发胖	易发胖	营养高	营养高	味苦	味苦	再次购买
4	同意	5	非常同意	4	同意	1
3	中立	3	中立	1	非常不同意	1
4	同意	3	中立	2	不同意	1
5	非常同意	3	中立	4	同意	1
3	中立	3	中立	2	不同意	1
2	不同意	4	同意	1	非常不同意	1
5	非常同意	2	不同意	4	同意	2
3	中立	3	中立	2	不同意	1
2	不同意	4	同意	1	非常不同意	1
4	同意	3	中立	2	不同意	1
4	同意	3	中立	4	同意	1
2	不同意	3	中立	1	非常不同意	1
2	不同意	4	同意	1	非常不同意	1
5	非常同意	2	不同意	4	同意	2
3	中立	3	中立	2	不同意	1
2	不同意	4	同意	1	非常不同意	1
1	非常不同意	4	同意	1	非常不同意	1
4	同意	3	中立	3	中立	1
4	同意	3	中立	4	同意	1
2	不同意	4	同意	1	非常不同意	1
2	不同意	4	同意	1	非常不同意	1
2	不同意	3	中立	1	非常不同意	1
5	非常同意	3	中立	4	同意	1
2	不同意	3	中立	1	非常不同意	1
3	中立	3	中立	2	不同意	1
4	同意	3	中立	2	不同意	1
2	不同意	3	中立	1	非常不同意	1

问卷 | 自动统计调查结果 | 编码设置

图 2.37　输入公式“=VLOOKUP(Y2,data,10,FALSE)”后的结果

第二十一步：在 AB 单元格中输入公式“=VLOOKUP(AA,data,11,FALSE)”，按 Enter 键后向下复制公式，其结果如图 2.38 所示。

=VLOOKUP(AA2,data,11,FALSE)

U	V	W	X	Y	Z	AA	AB
4	同意	5	非常同意	4	同意	1	是
3	中立	3	中立	1	非常不同意	1	是
4	同意	3	中立	2	不同意	1	是
5	非常同意	3	中立	4	同意	1	是
3	中立	3	中立	2	不同意	1	是
2	不同意	4	同意	1	非常不同意	1	是
5	非常同意	2	不同意	4	同意	2	否
3	中立	3	中立	2	不同意	1	是
2	不同意	4	同意	1	非常不同意	1	是
4	同意	3	中立	2	不同意	1	是
4	同意	3	中立	4	同意	1	是
2	不同意	3	中立	1	非常不同意	1	是
2	不同意	4	同意	1	非常不同意	1	是
5	非常同意	2	不同意	4	同意	2	否
3	中立	3	中立	2	不同意	1	是
2	不同意	4	同意	1	非常不同意	1	是
1	非常不同意	4	同意	1	非常不同意	1	是
4	同意	3	中立	3	中立	1	是
4	同意	3	中立	4	同意	1	是
2	不同意	4	同意	1	非常不同意	1	是
2	不同意	4	同意	1	非常不同意	1	是
2	不同意	3	中立	1	非常不同意	1	是
5	非常同意	3	中立	4	同意	1	是
2	不同意	3	中立	1	非常不同意	1	是
3	中立	3	中立	2	不同意	1	是
4	同意	3	中立	2	不同意	1	是
2	不同意	3	中立	1	非常不同意	1	是

卷 | 自动统计调查结果 | 编码设置

图 2.38　输入公式“=VLOOKUP(AA,data,11,FALSE)”后的结果

第二十二步：在“编码设置”工作表后面插入一个新的工作表，命名为“调查结果数据库”，在 A 列上输入“序号”；然后切换到“自动统计调查结果”工作表，按住 Ctrl 键，依次单击 B、D、F、H、J、L、N、P、R、T、V、X、Z、AB 列，选中替换后的调查结果，如图 2.39 所示。然后右击，在弹出的快捷菜单中选择“复制”选项将结果复制到“调查结果数据库”工作表中。选中 B1 单元格进行粘贴，其结果如图 2.40 所示。

	A	B	C	D	E	F	G	H	I	J
1	性别	性别	年龄	年龄	学历	学历	居住城市	居住城市	是否喝过啤酒	是否喝过啤酒
2	2	女	1	20-29	4	研究生及以上	2	吉林	1	是
3	1	男	2	30-39	3	本科	3	松原	1	是
4	1	男	2	30-39	3	本科	1	长春	1	是
5	2	女	1	20-29	3	本科	1	长春	1	是
6	2	女	1	20-29	2	大专	3	松原	1	是
7	1	男	2	30-39	4	研究生及以上	3	松原	1	是
8	2	女	1	20-29	2	大专	2	吉林	1	是
9	2	女	1	20-29	2	大专	1	长春	1	是
10	1	男	2	30-39	2	大专	1	长春	1	是
11	1	男	3	40-49	3	本科	2	吉林	1	是
12	2	女	2	30-39	3	本科	1	长春	1	是
13	1	男	3	40-49	1	高中及以下	3	松原	1	是
14	1	男	3	40-49	3	本科	4	白山	1	是
15	2	女	1	20-29	3	本科	2	吉林	1	是
16	2	女	1	20-29	3	本科	3	松原	1	是
17	1	男	3	40-49	4	研究生及以上	1	长春	1	是
18	1	男	2	30-39	1	高中及以下	2	吉林	1	是
19	1	男	4	50以上	3	本科	1	长春	1	是
20	2	女	2	30-39	3	本科	1	长春	1	是
21	1	男	2	30-39	3	本科	4	白山	1	是
22	1	男	1	20-29	3	本科	1	长春	1	是
23	1	男	2	30-39	2	大专	4	白山	1	是
24	2	女	2	30-39	3	本科	2	吉林	1	是

图 2.39　选中替换后的调查结果

	A	B	C	D	E	F	G	H
1	性别	年龄	学历	居住城市	是否喝过啤酒	最常喝的品牌	购买地点	饮用量（CC）
2	女	20-29	研究生及以上	吉林	是	青岛啤酒	大型超市	2000CC以上
3	男	30-39	本科	松原	是	蓝剑啤酒	专卖店	2000CC以上
4	男	30-39	本科	长春	是	蓝剑啤酒	专卖店	1000-2000CC
5	女	20-29	本科	长春	是	蓝剑啤酒	专卖店	500——999CC
6	女	20-29	大专	松原	是	金氏百啤酒	小卖部	1000-2000CC
7	男	30-39	研究生及以上	松原	是	青岛啤酒	大型超市	2000CC以上
8	女	20-29	大专	吉林	是	蓝剑啤酒	小卖部	500——999CC
9	女	20-29	大专	长春	是	蓝剑啤酒	专卖店	500——999CC
10	男	30-39	大专	长春	是	蓝剑啤酒	专卖店	2000CC以上
11	男	40-49	本科	吉林	是	青岛啤酒	专卖店	1000-2000CC
12	女	30-39	本科	长春	是	蓝剑啤酒	小卖部	500——999CC
13	男	40-49	高中及以下	松原	是	蓝剑啤酒	小卖部	1000-2000CC
14	男	40-49	本科	白山	是	蓝剑啤酒	专卖店	2000CC以上
15	女	20-29	本科	吉林	是	雪花啤酒	小卖部	500CC以下
16	女	20-29	本科	松原	是	青岛啤酒	大型超市	1000-2000CC
17	男	40-49	研究生及以上	长春	是	蓝剑啤酒	专卖店	2000CC以上
18	男	30-39	高中及以下	吉林	是	蓝剑啤酒	小卖部	2000CC以上
19	男	50以上	本科	长春	是	雪花啤酒	小卖部	500CC以下
20	女	30-39	本科	长春	是	蓝剑啤酒	小卖部	500CC以下
21	男	30-39	本科	白山	是	雪花啤酒	大型超市	2000CC以上
22	男	20-29	本科	长春	是	蓝剑啤酒	专卖店	2000CC以上
23	男	30-39	大专	白山	是	青岛啤酒	大型超市	2000CC以上
24	女	30-39	本科	吉林	是	蓝剑啤酒	小卖部	500CC以下
25	男	30-39	研究生及以上	松原	是	青岛啤酒	大型超市	2000CC以上
26	男	40-49	本科	白山	是	金氏百啤酒	小卖部	2000CC以上
27	女	20-29	研究生及以上	吉林	是	蓝剑啤酒	专卖店	1000-2000CC

调查问卷　自动统计调查结果　编码设置　调查结果数据库

图 2.40　粘贴后的结果

习　　题

一、填空题

1. 从使用者的角度看，统计数据的主要来源包括(　　)和(　　)两种渠道。
2. 某研究人员从公开的出版《中国统计年鉴》、《中国社会统计年鉴》获得数据，这是(　　)数据。

3. 调查数据常用的搜集方式有(　　)和(　　)。
4. 抽样调查具有(　　)、(　　)、(　　)和(　　)的特点。
5. 抽样调查的方式有很多，可以将这些不同的方式分为两类，即(　　)和(　　)。
6. 统计数据的误差主要包括(　　)和(　　)。

二、选择题

1. 我国第五次人口普查，是为了了解 2000 年 11 月 1 日零时人口的状况，某地区要求将调查单位资料于 2000 年 11 月 20 日前登记完毕，则普查的标准时间是(　　)。

A. 2000 年 11 月 20 日零时　　B. 2000 年 11 月 19 日 24 时
C. 2000 年 11 月 1 日 24 时　　D. 2000 年 10 月 30 日 24 时

2. 为了了解全国钢铁企业生产的基本情况，可对首钢、宝钢、鞍钢等几个大型企业进行调查，这种调查方式是(　　)。

A. 非全面调查　　B. 典型调查
C. 重点抽查　　D. 抽样调查

3. “你通常什么时候看电视？”此问题属于(　　)。

A. 事实性问题　　B. 假设性问题
C. 窘困性问题　　D. 以上都不是

4. 当需要把数值显示转换为具体内容时，在软件中通常使用的函数是(　　)。

A. SUM( )　　B. SUMIF( )
C. COUNTIF( )　　D. VLOOKUP( )

三、简答题

1. 简述问卷设计的原则。
2. 简述抽样调查的特点。

第3章 统计数据的整理与图形展示

教学目标

1. 掌握定性数据的整理方法和图形展示。
2. 掌握定量数据的整理方法和图形展示。
3. 了解合理使用统计图表。
4. 掌握数据的整理和图示展示的软件操作。

引入案例

某校在校大学生的平均每月消费结构的分析

在某大学随机抽取30名学生，调查他们的性别、家庭所在地、平均月生活费支出、平均每月购买衣物时所考虑的首要因素等，得到的数据见表3-1所示。

表3-1 平均每月消费结构的分析表

序号	性别	家庭所在地	平均月生活费/元	月平均衣物支出/元	买衣物首选因素
1	男	大型城市	800	200	价格
2	女	中小城市	600	180	款式
3	男	大型城市	1000	300	品牌
4	男	乡镇地区	700	40	价格
⋮	⋮	⋮	⋮	⋮	⋮
30	女	中小城市	500	50	价格

数据收集后，如果不对数据进行整理，是无法得出数据的规律性的。利用本章将要讲到的数据整理方法(利用 Excel 统计软件建立一个数据透视表，其中，建立数据透视表的步骤详看 3.4 节的内容)，可以轻易得出本次调查的分析结果，见表3-2所示。

表 3-2 调查分析结果

			家庭所在地			
性别	买衣服首选因素	数据	大型超市	乡镇地区	中小城市	总计
男	价格	求和项：平均月生活费/元	1 100	1 800	400	3 300
		求和项：月平均衣物支出/元	230	180	40	450
	款式	求和项：平均月衣物支出/元	500		3 000	3 500
		求和项：月平均衣物支出/元	150		800	950
	品牌	求和项：平均月生活费/元	1 000	800	1 600	3 400
		求和项：月平均衣物支出/元	300	240	480	1 020
男 求和项：平均月生活费/元			2 600	2 600	5 000	10 200
男 求和项：月平均衣物支出/元			680	420	1 320	2 420
女	价格	求和项：平均月生活费/元	700	400	2 600	3 700
		求和项：月平均衣物支出/元	230	120	465	815
	款式	求和项：平均月衣物支出/元	2 100		1 100	3 200
		求和项：月平均衣物支出/元	600		330	930
	品牌	求和项：平均月生活费/元	500	800		1 300
		求和项：月平均衣物支出/元	50	80		130
女 求和项：平均月生活费/元			3 300	1 200	3 700	8 200
女 求和项：月平均衣物支出/元			880	200	795	1 875

从本章开篇的案例可知，对数据进行分析，需要先对数据进行必要的整理。例如，对数据制作频数分布表、用图形进行展示等，以发现数据中的一些基本特征，为进一步分析提供思路。在对数据进行整理时，首先要弄清楚所面对的是什么类型的数据，因为不同类型的数据所采取的处理方法是不同的。

为了对数据进行分析，需要先对数据进行必要的整理。例如，对数据制作频数分布表、用图形进行展示等，以发现数据中的一些基本特征，为进一步分析提供思路。在对数据进行整理时，首先要弄清楚所面对的是什么类型的数据，因为不同类型的数据所采取的处理方法是不同的。

3.1 定性数据的整理与图形展示

3.1.1 定性数据的整理

根据前面的介绍，定性数据包括分类数据和顺序数据两种。对于定性数据的整理通常使用频数分布表。

定义 3.1　落在某一特定类别(或组)中的数据个数，称为频数。

定义 3.2　数据在各类别(或组)中的分配以表格形式展示，称为频数分布表。

通常在频数分布表中加入一列百分比，用百分比来反映样本(或总体)的构成或结构。

定义 3.3　一个样本(或总体)中各个部分的数据与全部数据之比，称为比例。

定义 3.4　将比例乘以 100 得到的数值称为百分比或百分数，用%表示。

百分比是一个更为标准化的数值，很多相对数都用百分比表示，当分子的数值很小而分母数值很大时，也可以用千分数(‰)来表示比例，如人口的出生率、死亡率、自然增长率等都用千分数来表示。

例如，一家饮料公司为研究自己产品的市场占有率，对随机抽取的一家超市进行调查。调查员在某天对 50 名顾客购买饮料的品牌进行了记录，如果顾客购买某一品牌的饮料，就

将这一饮料的品牌名字记录一次。最后得到的数据：购买可口可乐的顾客有 15 人，购买旭日升冰茶的顾客有 11 人，购买百事可乐的顾客有 9 人，购买汇源果汁的顾客有 6 人，购买露露的顾客有 9 人。这是一组分类数据，用频数分布表来整理，整理结果见表 3-3 所示。

表 3-3　不同饮料的频数分布表

饮料名称	频数	百分比/(%)
可口可乐	15	30
旭日升冰茶	11	22
百事可乐	9	18
汇源果汁	6	12
露露	9	18
合计	50	100

通常用 Excel 中的数据透视表来制作频数分布表，操作过程将在本章 3.4 节案例中详细介绍。

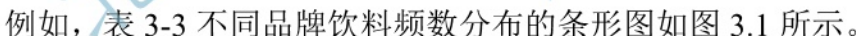
3.1.2　定性数据的图形展示

如果用图形来显示频数分布，就会更为形象和直观。一张好的统计图表，往往胜过冗长的文字表述。统计图的类型有很多，除了可以绘制二维平面图外，还可以绘制三维立体图。图形的制作均可由计算机来完成。

定性数据的图示方法包括条形图、帕累托图、对比条形图、饼图等；如果有两个总体或两个样本的分类相同且问题可比时，还可以绘制环形图。

1. 条形图

定义 3.5　用宽度相同的条形，用条形的高度或长短来表示数据频数多少的图形，称为条形图。



例如，表 3-3 不同品牌饮料频数分布的条形图如图 3.1 所示。

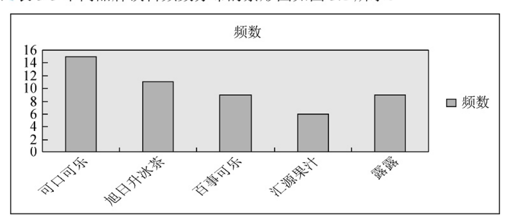


图 3.1　不同品牌饮料频数分布的条形图

2. 帕累托图

帕累托图是以意大利经济学家帕累托(Pareto)的名字来命名的。它也是条形图的一种特征情况。

定义 3.6　按各类别数据出现的频数多少经排序后绘制的条形图，称为帕累托图。

通过对条形图的排序，可以很容易地看出哪类数据出现得多，哪类数据出现得少。表 3-3 不同品牌饮料频数分布的帕累托图如图 3.2 所示。

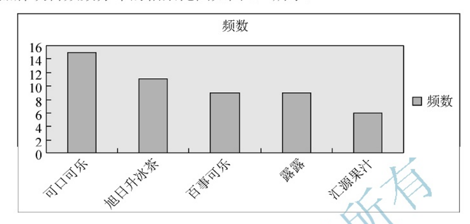


图 3.2　不同品牌饮料频数分布的帕累托图

3. 对比条形图

当分类变量在不同时间或不同空间上有多个取值时，为对比分类变量的取值，可以绘制对比条形图，了解数据在不同时间或不同空间上的差异或变化趋势。

例如，本月与上月的不同品牌饮料的市场调查结果见表 3-4 所示。

表 3-4　本月与上月的不同品牌饮料频数的市场调查数据

不同品牌饮料	本月	上月
可口可乐	15	12
旭日升冰茶	11	15
百事可乐	9	8
露露	6	7
汇源果汁	9	8
合计	50	50

利用 Excel 制作对比条形图，如图 3.3 所示。

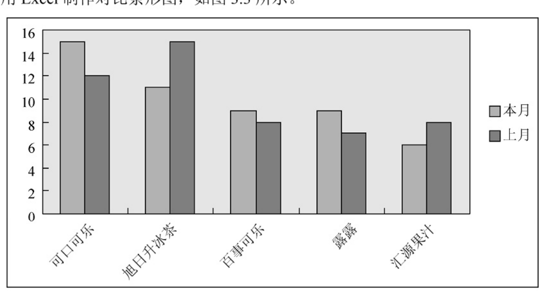


图 3.3　本月与上月不同品牌饮料的市场占有率对比

4. 饼图

定义 3.7　用圆形及圆内扇形的角度来表示数值大小的图形，称为饼图。它主要用于表示一个样本(或总体)中各组部分的数据占全部数据的比例。

饼图对于研究结构性问题十分有用。在绘制饼图时，样本中各部分所占的百分比可用圆内的各个扇形角度表示，即扇形的中心角度，是按各部分所占圆周的相应比例确定的。

例如，表 3-3 不同品牌饮料频数分布的饼图如图 3.4 所示。

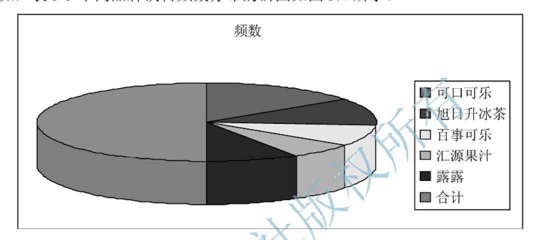


图 3.4　不同品牌饮料频数分布的饼图

3.2　定量数据的整理与图形展示

3.1 节介绍的定性数据的整理与图示方法，也可以对数值数据进行整理与显示。除此之外，数值数据还有一些特定的整理和图示方法，但它们并不适用于分类数据和顺序数据。

定量数据的整理有两种方法，即未分组和分组两种。对于不同整理的方式有不同的图形来展示。

3.2.1　未分组的定量数据的整理与图形展示

对于未分组的定量数据的整理，通常是对其进行简单的排序，通过对数据的排序，找出数据的规律性。未分组的数据通常用茎叶图和箱线图来展示。

1. 茎叶图

定义 3.8　由“茎”和“叶”两部分组成的、反映原始数据分布的图形，称为茎叶图。

茎叶图由“茎”和“叶”两部分构成，其图形是由数字组成的。通过茎叶图，可以看出数据的分布形状及数据的离散状况，如分布是否对称、数据是否集中、是否是离群点等。

绘制茎叶图的关键是设计好树茎。设计思路：树茎上长很多树叶，所以设计树茎时，要找出未分组数据的共同点为树茎，不同的为树叶。

例如，125、125、126、127、135、136、147、148 这组数据，可以用茎叶图表示，如图 3.5 所示。

茎	叶
12	5 5 6 7
13	5 6
14	7 8

图 3.5　茎叶图

茎叶图具有的特点：保留了原始数据的信息，通常适用于小批量数据。

2. 箱线图

定义 3.9　由一组数据的最大值、最小值、中位数和两个四分位数 5 个特征值绘制而成的、反映原始数据分布的图形，称为箱线图。

其中中位数和四分位数将在第 4 章学习，所以这里只是简单介绍一下箱线图如何制作。

箱线图由一个箱子和两条线段组成，如图 3.6 所示。其中左侧线段的起点由一组数据的最小值决定，右侧线段的终点由这组数据的最大值决定，箱子的左边和右边分别由这组数据的上下四分位数决定，箱子的里面有这组数据的中位数。

数据利用箱线图如何找出数据的规律性呢？简单来看一个例子：

例如从某大学经济学专业某一年级中随机抽取 3 名学生，3 名学生的这学期所有课程的成绩如下：

张小 1：英语 76 分；经济数学 65 分；西方经济学 93 分；市场营销学 74 分；会计学 68 分；政法经济学 70 分；统计学 55 分；计算机应用基础 85 分；

赵华 2：英语 90 分；经济数学 95 分；西方经济学 81 分；市场营销学 87 分；会计学 75 分；政法经济学 73 分；统计学 91 分；计算机应用基础 78 分；

王英 3：英语 97 分；经济数学 51 分；西方经济学 76 分；市场营销学 85 分；会计学 70 分；政法经济学 92 分；统计学 68 分；计算机应用基础 81 分；

根据中位数和上下四分位数的公式，可得出三名学生的箱线图，如图 3.7 所示。

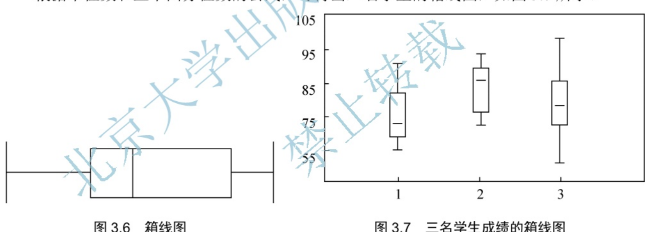


图 3.6　箱线图　　　　图 3.7　三名学生成绩的箱线图

从图 3.7 可以看出，在 3 名学生中，第 2 号学生(赵华)各科的平均考试成绩最高，而且各学科成绩之间的离散程度也较小；而第 1 号学生张小的平均考试成绩最低，而且各科考试成绩的离散程度也较大；各科考试成绩之间离散程度最大的是第 3 号学生王英。

3.2.2　分组的定量数据的整理与图形展示

1. 定量数据的分组

对数值数据进行整理的另一种方法，通常是对其进行分组。数据分组的主要目的是观察数据的分布特征。

定义 3.10　根据统计研究的需要，将原始数据按照某种标准分成不同的组别，称为数据分组。

其中，分组后的数据称为分组数据。数据经分组后再计算出各组中数据出现的频数，就形成了一张频数分布表。这里的分组是指组距分组。

定义 3.11　将全部变量值依次划分为若干个区间，并将这一区间的变量值作为一组，称为组距分组。其中组距分组又分为等距分组和不等距分组。

定义 3.12　各组组距相等的组距分组，称为等距分组。

例如，60～69 分为及格；70～79 分为中等；80～89 分为良好，这是一个等距分组。

定义 3.13　各组组距不相等的组距分组，称为不等距分组。

如，对人口年龄的分组，可根据人口成才的生理特定分成 0～6 岁(婴幼儿组)、7～17 岁(少年儿童组)、18～59 岁(中青年组)、60 岁以上(老年组)等。

为了让学生更好地掌握数值数据分组的过程，这里以例题的形式讲解如何进行分组，注意这里分组采用的是等距分组。

【例 3.1】长春市某年 1～2 月各天气温的记录数据见表 3-5 所示，对下面的数据进行适当的分组。

表 3-5　长春市某年 1～2 月各天气温　　单位：℃

−3	2	−4	−7	−11	−1	7	8	9	−6
−14	−18	−15	−9	−6	−1	0	5	−4	−9
−6	−8	−12	−16	−19	−15	−22	−25	−24	−19
−8	−6	−15	−11	−12	−19	−25	−24	−18	−17
−14	−22	−13	−9	−6	0	−1	5	−4	−9
−32	−4	−4	−16	−1	7	4	−6	5	4

定义 3.14　在组距分组中，一个组的最小值称为下限；一个组的最大值称为上限。

下面，结合例 3.1 说明分组的过程和频数分布表的编制过程。

(1) 对数据进行排序，排序后可得表 3-6。

表 3-6　长春市某年 1～2 月各天气温排序　　单位：℃

−32	−22	−17	−14	−11	−8	−6	−4	0	5
−25	−19	−16	−14	−9	−7	−6	−3	0	5
−25	−19	−16	−13	−9	−6	−4	−1	2	7
−24	−19	−15	−12	−9	−6	−4	−1	4	7
−24	−18	−15	−12	−9	−6	−4	−1	4	8
−22	−18	−15	−11	−8	−6	−4	−1	5	9

(2) 确定组数。确定组数是指将这组数据分成多少组。确定组数的规则条件有 3 个：

① 一般情况下，一组数据所分的组数 K 不应少于 5 组且不多于 15 组，即 $5 \leqslant K \leqslant 15$。

② 用斯特奇斯(Surges)提出的经验公式来确定组数 K：

$$K = 1 + \frac{\lg n}{\lg 2} \tag{3.1}$$

式中，n 为数据的个数。

对所得结果四舍五入取整数即为组数，则例 3.1 有 K=1+lg60÷lg2≈7，即应分为 7 组。当然，这只是一个经验公式。

③ 灵活确定组数。在实际应用中，可根据数据的多少和特点及分析的要求，参考以上两条灵活确定组数。由于这组数据不多，因此该例确定为 $K = 5$ 组。

(3) 确定各组的组距。

定义 3.15　一个组的上限与下限的差，称为组距。

组距可根据全部数据的最大值和最小值及所分的组数来确定，即

$$组距=(最大值－最小值)/组数 \tag{3.2}$$

在例 3.1 中，最大值为 9，最小值为−32，则组距=(9+32)/5=8.2。为便于计算，组距宜取 5 或 10 的倍数，因此组距可取 10。

(4) 分组并制作频数分布表。

采用组距分组时，需要遵循“不重不漏”的原则。“不重”是指一项数据只能分在其中的某一组，不能在其他组中重复出现；“不漏”是指组别能够包含所有数据，即在所分的全部组别中，每项数据都能分在其中的某一组，不能遗漏。为了解决“不漏”，所以第一组的下限应低于最小变量值，最后一组的上限应高于最大变量值。即第一组的下限就低于−32，最后一组的上限应高于 9。解决“不重”的问题，统计分组时习惯上规定“上组限不在内”，即当相邻两组的上下限重叠时，恰好等于某一组上限的变量值不算在本组内，而计算在下一组内。“不重不漏”用数学语言来表示就是分组后的变量值 x 满足 $a \leqslant x < b$ 。根据以上的原理，得出例 3.1 的分组后的频数分布表，见表 3-7 所示。

表 3-7　长春市某年 1～2 月各天气温分组后的频数分布表

按气温分组/℃	频数/天	频数/(%)
−40～−30	1	1.7
−30～−20	6	10
−20～−10	18	30
−10～0	25	41.7
0～10	10	16.6
合计	60	100

组距分组掩盖了各组内的数据分布状况，为反映各组数据的一般水平，通常用组中值作为该组数据的一个代表值。

定义 3.16　每一组的下限和上限之间的中点值，称为组中值，即组中值=(下限值＋上限值)/2。

使用组中值代表一组数据时有一个必要的假定条件，即各组数据在本组内呈均匀分布或在组中两侧呈对称分布。如果实际数据的分布不符合这一假定，那么用组中值作为一组数据的代表值会有一定的误差。

2. 分组数据的图形展示——直方图

通过数据分组后形成的频数分布表，可以初步看出数据分布的一些特征和规律。如果用图形来表示这一分布的结果，就会更为形象、直观。

通常使用直方图、折线图和曲线图来显示分组数据的频数分布特征，这里主要介绍直方图。

定义 3.17　用矩形的宽度和高度(即面积)来表示频数分布的图形，称为直方图。

在平面直角坐标中，用横轴表示数据分组，用纵轴表示频数或频率，那么各组与相应的频数就形成了一个矩形，即直方图。表 3-7 的直方图如图 3.8 所示。

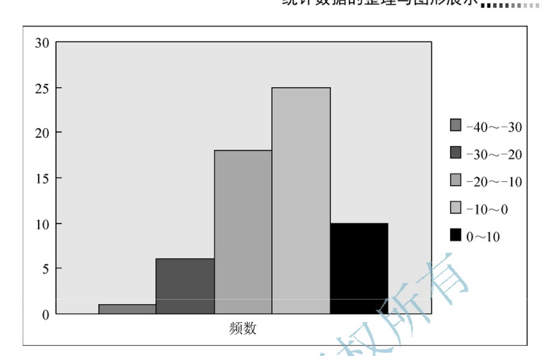


图 3.8　长春市某年 1～2 月各天气温(℃)的直方图

从图 3.8 可以直观地看出，对于温度的分布来说，其左边的尾部比右边的尾部长一些，略微有一些左偏分布。

直方图与条形图不同有以下 3 点：

(1) 条形图是用条形的宽度相同，用条形的长度表示各类别频数的多少，而直方图则是用面积表示各类别频数的多少。

(2) 从图 3.2 和图 3.8 中可以看出，直方图的各矩形通常是连续排列，而条形图则是分开排列。

(3) 条形图主要用于展示分类数据，而直方图则主要用于展示数值数据。

3. 时间序列数据的图形展示

如果数值数据是在不同时间上取得的，即时间序列数据，则可以绘制线图。

定义 3.18　线图是在平面坐标上用折线表现数据变化特征的图形。时间一般绘在横轴，观测值绘在纵轴。

线图主要用于显示时间序列数据，以反映事物发展变化的规律和趋势。

【例 3.2】 已知 2000—2011 年我国城乡居民家庭的人均收入数据见表 3-8 所示，试通过绘制线图来看我国城镇居民的家庭的人均收入发展变化。

表 3-8　2000—2011 年城乡居民家庭人均收入

单位：元

年份	城镇居民	农村居民
2000	6 280.0	2 253.4
2001	6 859.6	2 366.4
2002	7 702.8	2 475.6
2003	8 472.2	2 622.2
2004	9 061.22	3 582.42
2005	10 128.51	4 039.60
2006	11 320.77	4 631.21

续表

年份	城镇居民	农村居民
2007	12 719.19	5 025.08
2008	14 908.61	5 791.12
2009	17 067.78	6 700.69
2010	18 858.09	7 115.57
2011	21 033.42	8 119.51

解：绘制线图通常使用 Excel 软件来绘制，如图 3.9 所示。

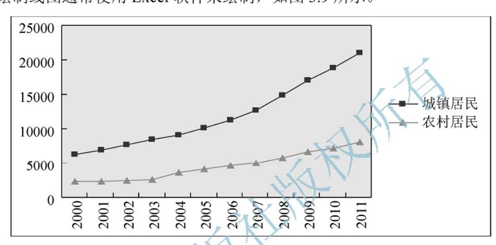


图 3.9　2000—2011 年城乡居民家庭人均收入线图

从图 3.9 中可以清楚地看出，城乡居民的家庭人均收入逐年提高，而且城镇居民的家庭人均收入高于农村。此外，自 2005 年后这种差距有扩大的趋势。

3.2.3　多维定量数据的整理与图形展示

上面介绍的一些图形描述的都是单变量数据。当有两个或两个以上变量时，利用一般的点图方法就很难描述了。为此，人们研究了多变量的图示方法，其中有散点图、三位散点图、雷达图等。

(1) 二维散点图。二维散点图是用二维坐标展示两个变量之间关系的一种图形。

(2) 三维散点图。当考察 3 个变量之间的关系时，二维散点图不再适用，这时可以绘制三位散点图和气泡图来展示 3 个变量之间的关系。

(3) 雷达图。雷达图是显示多个变量的常用图示方法。绘制雷达图的方法如下：设有 n 组样本 $S_1,S_2,\cdots,S_n$，每个样本测得 P 个变量 $X_1,X_2,\cdots,X_P$。我们在平面先做一个圆，然后将圆 P 等分(由变量的个数决定)，得到 P 个点，再将这 P 个点与圆心连线，得到 P 个辐射状的半径，这 P 个半径分别作为 P 个变量的坐标轴，每个变量值的大小由半径上的点到圆心的距离表示，再将同一样本的值在 P 个坐标上的点连线。这样一来，n 个样本形成的 n 个多边形就是一个雷达图。

3.3　统计表的使用

统计表和统计图是显示统计数据的两种方式。在日常生活中，当人们阅读报刊、看电视、查阅计算机网络时，都能看到大量的统计表格和统计图形。统计表把杂乱的数据有条

理地组织在一张简明的表格内，统计图把数据形象地显示出来。本节将介绍统计表的构成。

统计表一般由 4 个主要部分组成，即表头、行标题、列标题和数字资料。此外，必要时还可以在统计表的下方加上表外附加。某城市居民关注广告类型的频率分布见表 3-9 所示。

表 3-9　某城市居民关注广告类型的频数分布

广告类型	人数/人	频率
商品广告	112	56.0%
服务广告	51	25.5%
金融广告	9	4.5%
房地产广告	16	8.0%
招生招聘广告	10	5.0%
其他广告	2	1.0%
合计	200	100%

(资料来源：贾俊平. 统计学[M]. 2 版. 北京：清华大学出版社，2006.)

(1) 表头应放在表的上方，它所说明的是统计表的内容。表头一般应包括表号(表 3-9)、总标题(某城市居民关注广告类型的频数分布)和表中数据的单位等内容。总标题应简明确切地概括出统计表的内容，如果表中的全部数据都是同一计量单位，即可在表的右上角标明；若各变量的计量单位不同，则应放在每个变量后或单列出一列标明。

(2) 行标题和列标题通常安排在统计表的第一列和第一行，它所表示的主要是所研究问题的类别名称和变量名称。

(3) 表的其余部分是具体的数字资料；表外附加通常放在统计表的下方，主要包括数据来源、变量注释和必要的说明等内容。表中的数据一般是右对齐，有小数点时应以小数点对齐，而且小数点的位数应统一。对于没有数字的表格单元，一般用“—”表示，一张填好的统计表不应出现空白单元格。

(4) 在使用统计表时，必要时可在表的下方加上注释，特别要注明数据来源，以表示对他人劳动成果的尊重，这样也能方便读者查阅使用。

3.4　案例分析：啤酒市场的调查与分析及 Excel 上机应用——样本组成分析

3.4.1　性别结构的分析

由于抽样调查取得的样本中包含男性和女性,首先进行性别组成分析，这里使用筛选和图表功能来进行分析，其具体操作的过程如下。

第一步：插入一个新的工作表，命名为“样本组成分析”，在工作表中创建如图 3.10 所示的表格。

	A	B	C	D	E	F	G
1		一、性别组成分析					
2							
3		性别	频数				
4		男					
5		女					

调查问卷 / 自动录入调查结果 / 编码设置 / 调查结果数据库 / 样本组成分析

图 3.10 “样本组成分析”工作表

第二步：切换到“调查结果数据库”工作表中，选中第 1 行，单击“数据”→“筛选”按钮，此时第一行的单元格中将显示一个下拉按钮，如图 3.11 所示。

A1 性别

	A	B	C	D	E	F	G	H	I
1	性别	年龄	学历	居住城市	是否喝过啤	最常喝的品	购买地	饮用量（CC	增加气
2	女	20-29	研究生及以上	吉林	是	青岛啤酒	大型超市	2000CC以上	不同意
3	男	30-39	本科	松原	是	蓝剑啤酒	专卖店	2000CC以上	非常同意
4	男	30-39	本科	长春	是	蓝剑啤酒	专卖店	1000-2000CC	非常同意
5	女	20-29	本科	长春	是	蓝剑啤酒	专卖店	500——999CC	同意
6	女	20-29	大专	松原	是	金氏百啤酒	小卖部	1000-2000CC	非常同意
7	男	30-39	研究生及以上	松原	是	青岛啤酒	大型超市	2000CC以上	非常同意
8	女	20-29	大专	吉林	是	蓝剑啤酒	小卖部	500——999CC	同意
9	女	20-29	大专	长春	是	蓝剑啤酒	专卖店	500——999CC	非常同意
10	男	30-39	大专	长春	是	蓝剑啤酒	专卖店	2000CC以上	非常同意
11	男	40-49	本科	吉林	是	青岛啤酒	专卖店	1000-2000CC	非常同意
12	女	30-39	本科	长春	是	蓝剑啤酒	小卖部	500——999CC	同意
13	男	40-49	高中及以下	松原	是	蓝剑啤酒	小卖部	1000-2000CC	非常同意
14	男	40-49	本科	白山	是	蓝剑啤酒	专卖店	2000CC以上	非常同意
15	女	20-29	本科	吉林	是	雪花啤酒	小卖部	500CC以下	同意

图 3.11 数据筛选

第三步：单击第一行中的“性别”所在单元格的下拉按钮，在弹出的下拉列表中勾选“男”复选框，如图 3.12 所示。

	A	B	C	D	E	F
1	性别	年龄	学历	居住城市	是否喝过啤	最常喝的品
3	男	30-39	本科	松原	是	蓝剑啤酒
4	男	30-39	本科	长春	是	蓝剑啤酒
7	男	30-39	研究生及以上	松原	是	青岛啤酒
10	男	30-39	大专	长春	是	蓝剑啤酒
11	男	40-49	本科	吉林	是	青岛啤酒
13	男	40-49	高中及以下	松原	是	蓝剑啤酒
14	男	40-49	本科	白山	是	蓝剑啤酒
17	男	40-49	研究生及以上	长春	是	蓝剑啤酒
18	男	30-39	高中及以下	吉林	是	蓝剑啤酒
19	男	50以上	本科	长春	是	雪花啤酒
21	男	30-39	本科	白山	是	雪花啤酒
22	男	20-29	本科	长春	是	蓝剑啤酒
23	男	30-39	大专	白山	是	青岛啤酒
25	男	30-39	研究生及以上	松原	是	青岛啤酒
26	男	40-49	本科	白山	是	金氏百啤酒
28	男	30-39	大专	长春	是	蓝剑啤酒
29	男	20-29	本科	吉林	是	雪花啤酒
30	男	20-29	研究生及以上	长春	是	蓝剑啤酒
32						

图 3.12　设置筛选条件和筛选结果示意图

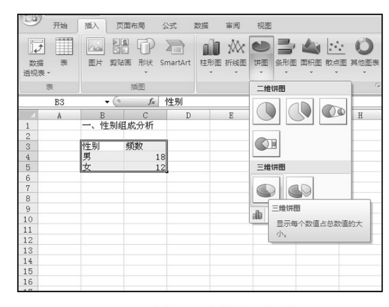
第四步：将上一步的结果填入在“样本组成分析”工作表中的表格中，结果如图 3.13 所示。

第五步：在“样本组成分析”工作表中，选中单元格区域 B3:C5，单击“插入”→“饼图”的下拉按钮，在弹出的下拉列表中选择“三维饼图”选项，如图 3.14 所示。

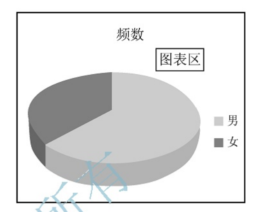
第六步：选中图形，右击，进行数据系列设置，得到的图形如图 3.15 所示。



图 3.13　将结果填入表格中



(a) 选择“三维饼图”选项



(b) 三维饼图

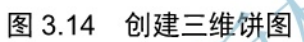


图 3.14 创建三维饼图

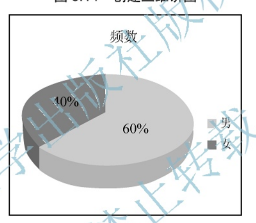


图 3.15 数据系列设置后的三维饼图

3.4.2 年龄结构的分析

下面对样本中受访者的年龄结构进行分析，这里利用数据透视图表来进行分析，具体的操作过程如下。

第一步：打开“样本组成分析”工作表，单击“插入”→“数据透视表”的下拉按钮，在弹出的下拉列表中选择“数据透视表”选项，弹出“创建数据透视表”对话框，如图 3.16 所示。

选择好区域及数据透视表的位置，如图 3.16 所示，单击“确定”按钮。

第二步：在“样本组成分析”工作表的右侧出现“数据透视表字段列表”任务窗格，如图 3.17 所示。

第三步：选中数据区域，单击“插入”→“柱形图”的下拉按钮，在弹出的下拉列表中选择“簇状柱形图”选项，如图 3.18 所示，得到如图 3.19 所示的结果。



图 3.16 “创建数据透视表”对话框



图 3.17 “数据透视表字段列表”任务窗格

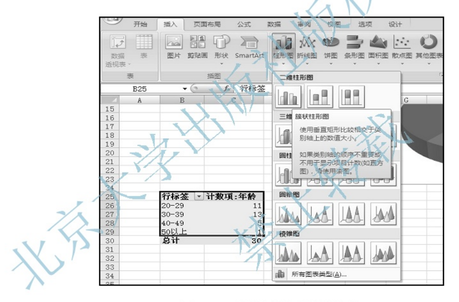


图 3.18 选择“簇状柱形图”

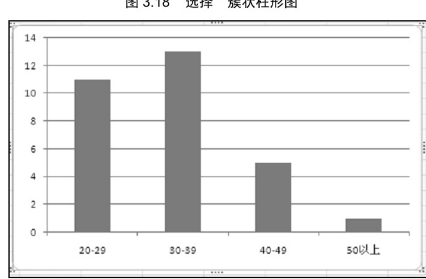


图 3.19 结果统计

习　题

一、填空题

1. 定性统计数据包括(　　)和(　　)两种类型。
2. 对于定性数据的整理我们通常使用的是(　　)。
3. 落在某一特定类别(或组)中的数据个数，称为(　　)。
4. 数据在各类别(或组)中的分配以表格形式展示，称为(　　)。
5. 用宽度相同的条形的高度或长短来表示数据多少的图形称为(　　)。
6. 各类别数据出现的频数多少经排序后绘制的图形称为(　　)。
7. 由“茎”和“叶”两部分组成的、反映原始数据分布的图形，称为(　　)。
8. 设计树茎时，对于未分组数据来说共同点为(　　)，不同的为(　　)。
9. 定量数据的整理有两种方法，分别是(　　)和(　　)两种。
10. 由一组数据的最大值、最小值、中位数和两个四分位数 5 个特征值绘制而成的、反映原始数据分布的图形，称为(　　)。
11. 将全部变量值依次划分为若干个区间，并将这一区间的变量值作为一组，称为(　　)。
12. 组距分组分为(　　)和(　　)。
13. 一个组的最小值称为(　　)；一个组的最大值称为(　　)。
14. 一个组的上限与下限的差，称为(　　)。
15. 采用组距分组时，需要遵循(　　)的原则。
16. 用矩形的宽度和高度(即面积)来表示频数分布的图形，称为(　　)。
17. 如果数值型数据是在不同时间上取得的，称为(　　)。
18. 用二维坐标展示两个变量之间关系的图形称为(　　)。
19. 统计表一般由 4 个主要部分组成，分别为(　　)、(　　)、(　　)、(　　)。
20. 表头一般应包括(　　)、(　　)和(　　)等内容。

二、单项选择题

1. 为对比分类变量的取值在不同时间或不同空间上的差异或变化趋势，可以绘制(　　)。
   A. 条形图　　B. 帕累托图　　C. 饼图　　D. 对比条形图
2. 对于未分组的数据通常用(　　)图形展示。
   A. 条形图　　B. 茎叶图　　C. 箱线图　　D. 直方图
3. 能够保留原始数据的信息的图形是(　　)。
   A. 饼图　　B. 茎叶图　　C. 箱线图　　D. 直方图
4. 下列不是构成箱线图的特征值的是(　　)。
   A. 最大值　　B. 最小值　　C. 众数　　D. 中位数
5. 将学生成绩分为 3 组：60～69 分为及格；70～79 分为中等；80～89 分为良好，这属于(　　)。
   A. 等距分组　　B. 不等距分组　　C. 组距分组　　D. 区间分组
6. 对人口年龄的分组，可分为以下 4 组：0～6 岁(婴幼儿组)、7～17 岁(少年儿童组)、18～59 岁(中青年组)、60 岁以上(老年组)，这种分组属于(　　)。
   A. 等距分组　　B. 不等距分组　　C. 组距分组　　D. 区间分组
7. 每一组的下限和上限之间的中点值，称为(　　)。
   A. 组中值　　B. 最小值　　C. 众数　　D. 中位数

8. 显示分组数据频数分布特征的图形是(　　)。
   A. 条型图　　B. 茎叶图　　C. 箱线图　　D. 直方图
9. 显示时间序列数据，主要使用(　　)。
   A. 条型图　　B. 茎叶图　　C. 线图　　D. 直方图
10. 考察3个变量之间的关系时，下列不适用的图形是(　　)。
   A. 二维散点图　　B. 三维散点图　　C. 气泡图　　D. 雷达图

三、练习题

1. 某银行为加强银行人员服务质量，特对服务质量设有5个等级供顾客选择，分别为：A——非常满意；B——满意；C——一般；D——不满意；E——非常不满意。为了解某一位职员服务的质量，从她所服务的顾客中随机抽取50人构成一个样本，调查结果如表3-10所示。

表3-10　调查结果

B	D	A	B	C	D	B	B	A	C
E	A	D	A	B	A	C	A	B	D
C	A	C	D	E	B	C	D	B	C
A	C	C	B	E	C	A	E	C	B
D	D	A	A	B	C	C	A	B	C

问题：

(1) 指出表3-10数据的类型。

(2) 用Excel制作一张频数分布表。

(3) 根据表3-10的数据绘制一张条形图。

2. 利用表3-11所示的数据绘制茎叶图和箱线图。

表3-11　调查数据

52	34	22	18	42	33	25	35	45	30
25	26	22	32	20	15	39	41	50	46
36	27	17	12	36	18	45	26	11	29
44	23	18	52	48	33	24	20	18	15

3. 某大型超市30天的销售额情况如表3-12所示。

表3-12　某大型超市30天的销售额

单位：万元

35	36	41	38	46	44	37	26	38	40
41	36	36	37	45	28	42	47	35	26
37	37	40	35	28	28	30	35	36	36

问题：

(1) 以组距为5对上面数据进行等距分组，并整理成频数分布表。

(2) 绘制直方图。

4. 某袋装洗衣液采用生产线自动装填，每袋容量大约为500mL，但由于某些原因，每袋容量不会恰好是500mL。表3-13是随机抽取的100袋产品的容量数据。

表 3-13　随机抽取的 100 袋产品的容量数据

500	516	528	485	509	491	484	505	518	519
506	515	512	482	491	508	490	492	507	501
508	529	494	481	495	485	506	461	535	465
468	510	493	497	474	458	498	498	496	498
506	492	491	527	499	482	498	500	510	522
494	490	536	489	496	471	473	529	508	527
488	489	483	485	502	521	498	513	486	502
501	681	518	507	483	517	523	512	483	492
493	497	494	481	521	520	487	479	495	491
513	499	525	526	507	501	503	496	517	488

问题：

(1) 利用计算机对表 3-13 的数据进行排序。

(2) 以组距为 10 为表 3-13 的数据进行等距分组，并整理成频数分布表。

(3) 绘制频数分布的直方图。

(4) 说明数据分布的特征。

5. A、B 两个班各有 35 名学生，期末经济学成绩的分布如表 3-14 所示。

表 3-14　学生的期末经济学成绩分布

考试成绩	人数	
	A 班	B 班
优	4	5
良	7	10
中	11	11
及格	9	6
不及格	4	3

问题：

(1) 根据表 3-14 的数据，画出两个班成绩的对比条形图和环形图。

(2) 比较两个班考试成绩的分布特点。

第 4 章　统计数据的指标度量

教学目标

1. 掌握数据集中趋势的度量。
2. 掌握数据离散程度的度量。
3. 了解数据的偏态和峰态的度量。
4. 掌握描述统计指标的软件操作过程。

引入案例

哪名运动员的发挥更稳定?

在奥运会女子 10 米气手枪比赛中，每个运动员首先进行每组 10 枪共 4 组的预赛，然后根据预赛总成绩确定进入决赛的 8 名动动员。决赛时 8 名运动员再进行 10 枪射击，预赛成绩加上决赛成绩确定最后的名次。在 2008 年 8 月 10 日举办的第 29 届北京奥运会女子 10 米气手枪决赛中，进入决赛的 8 名运动员的预赛成绩和最后 10 枪的决赛成绩见表 4-1 所示。

表 4-1　8 名运动员决赛成绩

姓名	国家	预赛成绩	决赛 10 枪成绩/环									
纳塔利娅	俄罗斯	391	10	8.5	10	10.2	10.6	10.5	9.8	9.7	9.5	9.3
郭文珺	中国	390	10	10.5	10.4	10.4	10.1	10.3	9.4	10.7	10.8	9.7
卓格巴德拉赫	蒙古	387	9.3	10	8.7	8.3	9.2	9.5	8.5	10.7	9.2	9.2
妮诺	格鲁吉亚	386	9.8	10.3	10	9.5	10.2	10.7	10.4	10.6	9.1	10.8
维多利亚	白俄罗斯	384	9.3	9.4	10.4	10.1	10.2	10.5	9.2	10.5	9.8	8.6
莱万多夫斯卡	波兰	384	8.1	10.3	9.2	9.9	9.8	10.4	9.9	9.4	10.7	9.6
亚斯娜	塞尔维亚	384	10.2	9.6	9.9	9.9	9.3	9.1	9.7	10	9.3	9.9
米拉	芬兰	384	8.7	9.3	9.2	10.3	9.8	10	9.7	9.9	9.9	9.7

最后得出的结论是，塞尔维亚的运动员发挥是最稳定的，其次是中国运动员郭文珺，试问得出结论的依据是什么呢?

4.1　集中趋势的指标

集中趋势是指一组数据向某一中心值靠拢的程度，它反映了一组数据中心点的位置所在，即集中趋势是找出一组数据的代表值。前面已经介绍过数据的分类，其中以计量尺度

不同，可以把数据分为分类型数据、顺序型数据和数值型数据，且 3 种不同类型数据，是从低层次测量数据到高层次测量数据，本节选用这 3 种不同类型数据分别介绍不同数据集中趋势的指标。需要强调的是，低层次数据的集中趋势指标适用于高层次测量数据，而反过来，高层次测量数据的集中趋势指标不适用于低层次的测量数据。

4.1.1 分类数据——众数

1. 众数的定义

定义 4.1 一组数据中出现频数最多的变量值，称为众数，用 M_0 表示。

众数主要用于测量分类型数据的集中趋势，同时也适用于测量顺序数据和数值型数据。

【例 4.1】 某研究人员记录的被调查者的性别见表 4-2 所示，试计算这组数据的众数。

表 4-2 被调查者的性别数据

男	女	女	女	男
男	女	女	男	女
女	女	女	女	男
女	女	女	男	男

解：这里的变量为性别，它是分类变量，因而不同的性别就是不同的变量值，所以这里只有“男”和“女”两个变量值。在所有 20 位被调查者中，“男”变量值有 7 位，所以它的频数为 7，“女”变量值有 13 位，所以它的频数为 13，相对“男”变量值，“女”变量值的频数最多，所以此题的众数为“女”。

【例 4.2】 金融学院 2010 级学生在本学期统计学成绩见表 4-3 所示，试计算这组数据的众数。

表 4-3 统计学成绩数据

成绩等级	人数
优秀	10
良好	23
中等	23
及格	5
不及格	2

解：这里的变量为“成绩等级”，其变量值为“优秀”、“良好”、“中等”、“及格”和“不及格”。从表中看，“良好”和“中等”两个变量值的频数为 23，比其他的变量值都多，所以众数为“良好”和“中等”。

【例 4.3】 在某城市中随机抽取了 9 个家庭，调查得到每个家庭的人均月消费数据如下(单位：元)，计算人均月消费的众数。

450 560 900 1080 570 810 710 720 460

解：这里的变量是数值变量，变量值为 450、560、900、1 080、570、810、710、720、460。从数值看，每一个变量值的频数都为 1，因此这组数据没有最多的频数，所以此组数据无众数。

2. 众数的特点

由以上的例题可以总结出众数具有以下特点。

(1) 众数是一个位置代表值，它不受极端值的影响。

例如，一组数据 750、850、960、1 080、1 080、1 250、1 630、2 000，此组数据的众数 M_0=1 080，当这组数据中最小值 750 变为 550 时，众数还是 M_0=1 080。所以它不受极端值的影响。

(2) 众数也可能不存在；如果存在，并不是唯一的，可能有一个，也可能有两个，甚至更多。

例如，例 4.1 有一个众数，例 4.2 有两个众数，例 4.3 无众数。

(3) 一组数据的众数代表数据是否有明显的集中趋势或最高峰点，所以众数的示意图如图 4.1 所示。

3. Excel 中的众数计算函数

利用 Excel 中的 MODE 函数可以计算出一组数值数据的众数。其中语法为 MODE (number1,number2,⋯)，如果一组数据中不含有众数，则 MODE 函数返回错误值 N/A。

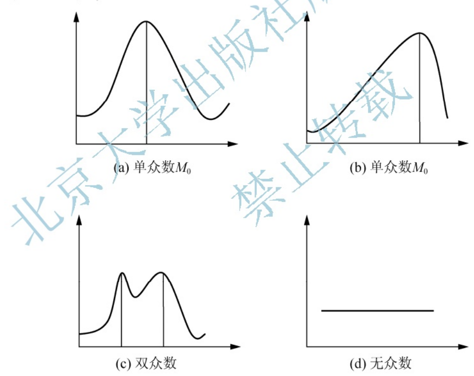


图 4.1 众数示意图

4.1.2 顺序数据：中位数和四分位数

分位数是指在一组排序好的数据中，处于某个位置上的数据，就是相应的分位数，包括中位数、四分位数、十分位数和百分位数等。以下主要介中位数和四分位数。

1. 中位数

1) 中位数的定义

定义 4.2　一组数据按顺序排列后处于中间位置上的变量值，称为中位数，用 M_e 表示。

由中位数的定义可知，中位数是把一组数据分成相等的两部分，每部分的数据包含整组数据的 50%，其中一部分数据比中位数小，另一部分数据比中位数大。中位数主要是测量顺序数据的集中趋势，它也可以测量数值数据的集中趋势，但不可以测量分类数据的集中趋势。

2) 中位数的计算公式

根据中位数的定义，计算中位数，首先要对数据进行排序，然后确定中位数的位置，这个位置上所对应的变量值就是中位数。那么中位数的位置确定的公式为

$$\text{中位数位置}=\frac{n+1}{2} \tag{4.1}$$

下面来看几个例题，由例题推出中位数的计算公式。

【例 4.4】 根据表 4-2 的数据，计算数据的中位数。

解：这是一个顺序数据，变量为“成绩等级”，其中“优秀”、“良好”、“中等”、“及格”和“不及格”是变量值。由于变量值本身就是排序的，根据中位数的定义得知处于中间位置的为

$$\text{中位数的位置}=\frac{n+1}{2}=\frac{(10+23+23+5+2)+1}{2}=32$$

即此组数据是处于第 32 个位置的所对应的变量值，位于“良好”这个区间，所以 $M_e=$ 良好。

【例 4.5】 计算例 4.3 数据的中位数。

450　560　900　1 080　570　810　710　720　460

解：这是一个数值数据，变量为人均月消费，要想计算中位数，首先把数据进行排序，如下：

450　460　560　570　[710]　720　810　900　1 080

共 9 个数据，中位数的位置 $=\frac{n+1}{2}=5$，即为 710，此时中位数 $M_e=x_{\frac{n+1}{2}}$

【例 4.6】 在例 4.3 的数据组中加入一个数据 1 200 后，计算新数据组的中位数。

解：加入一个数据 1 200 后，新数据排序后如下：

450　460　560　570　[710]　[720]　810　900　1 080　1 200

这时新数据组共有 10 个数据，则中位数的位置 $=\frac{n+1}{2}=5.5$，即中位数处于第 5 个位置(对应的变量值为 710)和第 6 个位置(对应的变量值为 720)中间的数据，则为两个数据的加和除 2，即 $M_e=\frac{1}{2}(x_{\frac{n}{2}}+x_{\frac{n}{2}+1})=715$。

从上面的 3 个例题可以得出中位数的计算公式为

$$M_e = \begin{cases} x_{\frac{n+1}{2}} & n\text{为奇数} \\ \frac{1}{2}(x_{\frac{n}{2}} + x_{\frac{n}{2}+1}) & n\text{为偶数} \end{cases} \tag{4.2}$$

3) 中位数的特点

中位数是主要测量顺序数据的集中趋势指标，是一个位置代表值，所以中位数不受数据组的极端值影响。

4) Excel 中的中位数的计算函数

利用 Excel 中的 MEDIAN 函数可以计算出一组数值数据的中位数。其语法为 MEDIAN(number1,number2,…)。

2. 四分位数

1) 四分位数的定义及计算公式

定义 4.3　一组数据排序后处于 25%和 75%位置上的值，称为四分位数。

四分位数是把一组数据四等分，一组数据四等分，需要 3 个点，分别为此组数据的中位数和被分成的两部分数据各自的中位数，即整组数据的 25%、50%和 75%三个位置上的变量值。其中 50%位置上的变量值称为中位数，剩余下的两个数就称为四分位数，处于 25%位置上的变量值是下四分位数，用 Q_L 表示；处于 75%位置上的变量值是上四分位数，用 Q_U 表示。

四分位数的确定与中位数确定不同的是，四分位数位置的确定方法有很多种，每种方法的结果有一定的差异，但差异不是很大，本书四分位数的位置确定公式为

$$Q_L\text{ 位置} = \frac{n}{4} \qquad Q_U\text{ 位置} = \frac{3n}{4} \tag{4.3}$$

2) Excel 中的四分位数的计算函数

利用 Excel 中的 QUARTILE 函数可以计算出一组数值数据的四分位数。其中语法为 QUARTILE(array quart)，array 为需要求得四分位数值数据的单元格区域，quart 有 4 种取值，分别为 0、1、2、3 和 4，不同的取值，返回不同的四分位值。当取值为 0 时，返回这组数据的最小值；当取值为 1 时，返回这组数据的下四分位数；当取值为 2 时，返回的是这组数据的中位数；当取值为 3 时，返回的是这组数据的上四分位数；当取值为 4 时，返回这组数据的最大值。

【例 4.7】 根据例 4.3 中 9 个家庭的消费调查数据，计算人均月消费的四分位数。

解：排序后的数据为

450　460　560　570　710　720　810　900　1 080

共 9 个数据，则 Q_L 位置 $= \frac{n}{4}$ =2.25，即处于第 2 个位置上(变量值为 460)和第 3 个位置上(变量值为 560)间，则

$$Q_L = 460 + 0.25 \times (560 - 460) = 485$$

Q_U 位置 $= \frac{3n}{4}$ =6.75，即处于第 6 个位置上(变量值为 720)和第 7 个位置上(变量值为 810)间，则

$$Q_U = 720 + 0.75 \times (810 - 720) = 787.5$$

4.1.3 数值数据——平均数

定义 4.4 一组数据相加后除以数据的个数而得到的结果，称为平均数。

平均数是测量数值数据的集中趋势，不适用于分类数据和顺序数据。根据第 3 章的内容，数值数据的整理有两种：一是未分组数值数据；二是分组数值数据，所以针对这两种不同整理方式的数值数据有两个不同的平均数计算形式和计算公式。

1. 简单平均数

1) 简单平均数的定义及计算公式

简单平均数主要是测量未分组数值数据的集中趋势。

定义 4.5 未经分组的一组样本数据为 $x_1, x_2, \cdots, x_n$，样本数据的个数为 n，则样本平均数为 $\bar{x}$，称为简单平均数。

简单平均数的计算公式为

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n} \tag{4.4}$$

【例 4.8】 根据例 4.3 中的 9 个家庭的消费调查数据，计算人均月消费的简单平均数。

解：根据式(4.4)，有

$$\begin{aligned}\bar{x} &= \frac{x_1 + x_2 + \cdots + x_n}{n} \\ &= \frac{450+460+560+570+710+720+810+900+1\,080}{9} \\ &= \frac{6\,260}{9} = 695.6\end{aligned}$$

2) 简单平均数的特点

简单平均数的数值受这组数据的极端值影响，即改变这组数据的极大值和极小值时，此组数据的平均数也随之改变。

2. 加权平均数

1) 加权平均数的定义及计算公式

加权平均数主要是测量分组数值数据的集中趋势。

定义 4.6 将原始数据分成 K 组，各组的数据用各自组中值来表示，各组变量值出现的个数用各自的频数表示，这样的数据平均数称为加权平均数。

设原始数据分 k 组，各组的组中值分别为 $M_1, M_2, \cdots, M_k$，各组变量值出现的频数为 $f_1, f_2, \cdots, f_k$，则样本平均数的计算公式为

$$\bar{x} = \frac{M_1 f_1 + M_2 f_2 + \cdots + M_k f_k}{f_1 + f_2 + \cdots + f_k} = \frac{\sum_{i=1}^{k} M_i f_i}{n} \tag{4.5}$$

式中，$n = f_1 + f_2 + \cdots + f_k$，即样本容量。

【例 4.9】某企业职工年收入统计资料见表 4-4 所示，试计算这个企业职工平均年收入。

表 4-4　某企业职工年收入数据

年收入/万元	职工人数/人
2.0～3.0	6
3.0～4.0	10
4.0～5.0	18
5.0～6.0	12
6.0～7.0	7
合计	53

解：这是一组分组数值数据，计算平均数使用加权平均数。根据式(4.5)有

$$\begin{aligned}\bar{x}&=\frac{M_1f_1+M_2f_2+\cdots+M_kf_k}{f_1+f_2+\cdots+f_k}\\&=\frac{6\times2.5+10\times3.5+18\times4.5+12\times5.5+7\times6.5}{53}\\&=4.58\end{aligned}$$

其中每组的组中值分别为 2.5、3.5、4.5、5.5、6.5，各组的频数分别为 6、10、18、12 和 7。

从加权平均数的公式得知，用各组的组中值来替代各组的实际变量值，如果各组数据在组内是均匀分布的，则计算结果还是比较准确的；如果各组数据在组内是非均匀分布的，则误差较大。所以一般情况下，在各组数据均匀分布的前提下，使用加权平均数来计算平均数。

2) 加权平均数的特点

从加权平均数的定义可以得知，其数值的大小不仅受各组组中值的影响，而且受各组变量值频数的影响。即当一组频数较大时，意味着这组数据的个数较多，则这组数据的组中值对平均数的影响较大。

3. Excel 中的加权平均数的计算函数

利用 Excel 中的 AVERAGE 函数可以计算出一组数值数据的平均数。其语法为 AVERAGE(number1,number2,…)，返回其参数的算术平均值，参数可以是数值或包含数值的名称、数组或引用。

4.1.4　众数、中位数和平均数的关系

众数、中位数和平均数都是测量数值数据的集中趋势度量指标，则三者存在的关系如下。

(1) 当一组数据是对称分布时，众数 M_0、中位数 M_e 和平均数 $\bar{x}$ 必定相等，即 $M_0=M_e=\bar{x}$，如图 4.2(a)所示。

(2) 当一组数据是左偏时，众数 M_0、中位数 M_e 和平均数 $\bar{x}$ 的关系为 $\bar{x}<M_e<M_0$，如图 4.2(b)所示。

原因：左偏分布，说明数据存在极小值，而中位数和众数不受极端值的影响，而平均值受极端值影响，所以平均数向极小值一方发展。

(3) 当一组数据是右偏时，那么众数 M_0、中位数 M_e 和平均数 $\bar{x}$ 的关系为 $M_0<M_e<\bar{x}$，

如图 4.2(c)所示。

原因：右偏分布，说明数据存在极大值，而中位数和众数不受极端值的影响，而平均值受极端值影响，所以平均数向极大值一方发展。

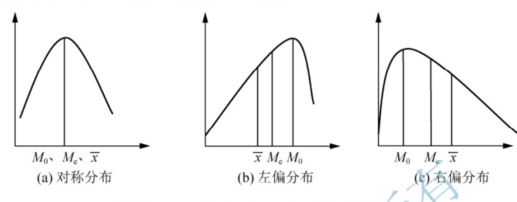


图 4.2　不同分布的众数、中位数和平均数的关系

根据众数、中位数和平均数三者存在的关系，给出一组数据，可以计算出这组数据的众数、中位数和平均数，利用三者的关系，初步判断出这组数据的分布情况。

4.1.5　众数、中位数和平均数应用的注意事项

(1) 众数可以测量分类数据、顺序数据和数值数据，但主要适用于测量分类数据的集中趋势度量。如果一组数据量较多时才有意义，如果一组数据量较少时，不宜采用众数。

(2) 中位数是一组数据位置上的代表值，它不受极端值的影响。当一组数据的分布偏斜程度较大时，中位数是集中趋势测量的一个很好选择指标。但它也是最主要测量顺序数据的。

(3) 平均数是针对数值数据集中趋势的指标，它不可以测量其他数据。但对于数值数据的指标却有 3 个，选择使用哪个指标是非常重要的。当一组数据是非对称分布时，如果选择平均数测量数值数据，这时平均数的代表性就较差，因为平均数受极端值的影响，这时要考虑使用众数或中位数；如果一组数据是对称分布，数值数据采用平均数作为它的集中趋势测量指标。

4.2　离散程度的绝对指标

集中趋势是选出一组数据的代表值，选出代表值后，也要对代表值的代表性进行评价，对一组数据各变量值向代表值靠拢的程度，或各变量值之间的差异状况进行评价，即数据离散程度的度量。

离散程度是指一组数据中各变量值远离其中心值(代表值)的程度。数据的离散程度越大，集中趋势选出的代表值的代表性就越差；离散程度越小，代表值的代表性就越强。

本节同样按照数据的计量尺度不同，即分类数据、顺序数据和数值数据(从低到高)顺序来讲解。

4.2.1　分类数据——异众比率

定义 4.7　非众数组的频数占总频数的比率，称为异众比率，用 V_r 表示。

根据定义，可得知异众比率的计算公式为

$$V_r = \frac{\sum f_i - f_m}{\sum f_i} \tag{4.6}$$

式中，$\sum f_i$ 为变量值的总频数；f_m 为众数组的频数。

【例 4.10】 一家市场调查公司为研究不同饮料的市场占有率，对随机抽取的一家超市进行调查。调查人员在某天 50 记录名顾客购买的饮料，经统计得到表 4-5，试计算异众比率。

表 4-5　不同饮料的频数分布表 1

不同饮料	频数
碳酸饮料	15
冰红茶	11
冰糖雪梨	9
果汁	6
矿泉水	9
合计	50

解：由表 4-5 可知，这是分类数据，且众数 M_0=碳酸饮料，其他变量值为非众数。

$$V_r = \frac{\sum f_i - f_m}{\sum f_i} = \frac{50-15}{50} = 70\%$$

这说明，在所调查的 50 人中，有 70%的人购买的不是碳酸饮料，只有 30%的人购买了碳酸饮料。

【例 4.11】 将表 4-5 中的数据更改为表 4-6 的数据，试计算异众比率。

表 4-6　不同饮料的频数分布表 2

不同饮料	频数
碳酸饮料	35
冰红茶	5
冰糖雪梨	5
果汁	4
矿泉水	1
合计	50

解：由表 4-6 中可知，众数还是 M_0=碳酸饮料，其他变量值为非众数。

$$V_r = \frac{\sum f_i - f_m}{\sum f_i} = \frac{50-35}{50} = 30\%$$

这说明，在所调查的 50 人，有 30%的人购买的不是碳酸饮料，70%的人购买了碳酸饮料。

比较例 4.10 和例 4.11 的结果，可以看出，例 4.11 的异众比率小于例 4.10 的，而例 4.11 的异众比率说明调查的 50 人，只有 30%没有购买碳酸饮料(众数)，说明在例 4.11 中，众数

比例 4.10 中的代表性强。由此，得出以下结论。

异众比率越大，说明众数的代表性就越差；异众比率越小，说明众数的代表性就越强。

异众比率主要测量分类数据的离散程度，同时也可以测量顺序数据和数值数据的离散程度。

4.2.2 顺序数据——四分位差

定义 4.8 上四分位数与下四分位数之差，也称内距或四分间距，称为四分位差，用 Q_d 表示。

根据四分位差的定义，得出四分位差的公式：

$$Q_d = Q_U - Q_L \tag{4.7}$$

由 4.1 节的内容知道，下、上四分位数分别是处于一组数据 25%位置和 75%位置上的变量值，也就是说四分位差包含了一组数据中间的 50%数据，是反映中间 50%数据的离散程度。如果四分位差的值越小，意味着中间 50%的数据越向中位数靠拢，即中位数的代表性就越强；如果四分位差的值越大，意味着中间 50%的数据越分散，中位数的代表性就越弱。

【例 4.12】 根据例 4.7 的计算结果，计算家庭人均月消费的四分位差。

解：根据例 4.7 的计算结果可知：

$$Q_U = 787.5 \qquad Q_L = 485$$

所以四分位差为

$$Q_d = Q_U - Q_L = 785.5 - 485 = 300.5$$

四分位差主要用于测量顺序数据，同时也可以测量数值数据离散程度，但不适用于测量分类数据的离散。

4.2.3 数值数据——方差和标准差

定义 4.9 各变量值与其平均数离差平方的平均数，称为方差。总体的方差用 σ^2 表示，样本方差用 s^2 表示。

定义 4.10 方差的平方根称为标准差。

标准差是最常用最基本的一种标志变异指标。总体的标准差用 σ 表示，样本标准差用 s 表示。

1. 样本方差和标准差的计算

根据样本数据分为分组数据和未分组数据，有两种样本方差和标准差的计算公式。

1) 未分组数据

根据方差的定义，样本方差计算公式为

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1} \tag{4.8}$$

2) 分组数据

分组数据，各变量值用所在组的组中值来替代，所以其计算公式为

$$s^2 = \frac{\sum_{i=1}^{k} f_i (M_i - \overline{x})^2}{n-1} \tag{4.9}$$

3) 自由度

以上两个公式的分母都为$n-1$，即样本数据的个数减 1，称之为自由度。自由度是指在一组数据有一个附加的约束时，自由取值的变量个数。

例如，$x_1 + x_2 = 2\overline{x}$，其中$\overline{x} = 1$，即$x_1 + x_2 = 2$，两个变量相加和为 2，这时两个变量度自由取值的只有一个变量，1 为此数据的自由度。

再如，$x_1 + x_2 + x_3 = 3\overline{x}$，$\overline{x}$再固定，3 个变量自由取值的只有两个，2 为自由度。依次类推，在样本方差公式中，有$\overline{x}$这个附加的约束，所以有$x_1 + \cdots + x_n = n\overline{x}$，$n$个变量中自由取值的有$n-1$个，$n-1$为自由度。

4) 样本标准差的计算

根据标准差的定义，有其计算公式为

$$s = \sqrt{s^2} \tag{4.10}$$

2. 总体方差和标准差的计算

1) 未分组总体方差的计算

未分组总体方差的计算公式为

$$\sigma^2 = \frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N} \tag{4.11}$$

2) 分组总体方差的计算

分组总体方差的计算公式为

$$\sigma^2 = \frac{\sum_{i=1}^{k} f_i (x_i - \mu)^2}{N} \tag{4.12}$$

3) 总体的标准差计算

$$\sigma = \sqrt{\sigma^2} \tag{4.13}$$

3. Excel 中标准差和方差的计算函数

1) 标准差的计算函数

利用 Excel 中的 STDEV 函数可以计算出一组数值数据的标准差。其中语法为 STDEV(number1,number2,…)，返回其参数的标准差，参数可以是数值或包含数值的名称、数组或引用。

2) 方差的计算函数

利用 Excel 中的 VAR 函数可以计算出一组数值数据的方差。其语法为 VAR(number1,number2,…)，返回其参数的方差，参数可以是数值或包含数值的名称、数组或引用。

【例 4.13】 根据表 4-3 的数据，计算该企业职工年收入的标准差。

解：根据样本方差公式有

$$s^2 = \frac{\sum_{i=1}^{k} f_i (M_i - \bar{x})^2}{n-1}$$

$$= \frac{6 \times (2.5-4.6)^2 + 10 \times (3.5-4.6)^2 + 18 \times (4.5-4.6)^2 + 12 \times (5.5-4.6)^2 + 7 \times (6.5-4.6)^2}{53-1}$$

$=1.42$

$$s = \sqrt{s^2} = \sqrt{1.42} = 1.19$$

4.2.4 相对离散程度——离散系数

以上介绍了数值数据离散程度的测量指标，即方差和标准差，方差和标准差反映的是各变量值变异程度的绝对指标，其数值的大小受各变量值变动程度的影响，且也受平均水平的影响。例如，测量 20 人身高的离散程度，当不同研究人员采用不同单位时，即研究人员采用 m 为单位，得出一个标准差数值，另一个研究人员采用 cm 为单位，得出另一个标准差数值，两个数值不相等。为了消除计量单位不同或平均水平高低不等的影响，采用反映离散程度的相对指标，即离散系数。

定义 4.11　一组数据的标准差与其相应的平均数之比，称为离散系数，用 V_s 表示。

其计算公式为

$$V_s = \frac{s}{\bar{x}} \tag{4.14}$$

离散系数是离散程度测量的相对指标，可以应用于比较不同样本数据的离散程度。

离散系数大的，说明数据的离散程度越大，即偏离代表值；离散系数小的，说明数据的离散程度越小，即一组数据向代表值靠拢。

4.3 数据的相对位置测量——标准分数

在生活中，人们经常要测量数据的相对位置，即测量某个数据在一组数据中的位置。

例如，某个同学在期末考试中货币银行学取得 89 分，统计学取得 75 分，试问他哪科成绩较好？这时我们不能用分数的绝对值来进行衡量，因为会存在一种可能，货币银行学全班取的成绩都较高或货币银行学的试卷较简单，而统计学试卷较难，所以不能用绝对值衡量他哪科学得好。这时我们可以采用相对指标，即计算这两科分数在全班分数数据中的位置。

定义 4.12　变量值与其平均数的离差除以标准差后的值，称为标准分数，用 Z 表示。

根据标准分数定义，其计算公式为

$$Z = \frac{x_i - \bar{x}}{s} \tag{4.15}$$

标准分数是测量各变量值在所在的数据中的位置。例如，如果某个变量值的标准分数为 -1 时，说明其变量值低于平均值，且低于一个标准差；如果某个变量值的标准分数为 1 时，说明其变量值高于平均值，且高出一个标准差。

【例 4.14】 某个同学在期末考试中货币银行学取得 89 分，全班同学的货币银行学均值为 75 分，标准差为 7 分；统计学取得 75 分，全班同学统计学均值为 60 分，标准差为 5 分。试问，这名同学哪科学得好？

解：比较该名学生的哪科学得好，只需测量出他每科成绩在全班的位置，即相对位置。所以有

货币银行学的位置：

$$Z = \frac{x_i - \bar{x}}{s} = \frac{89 - 75}{7} = 2$$

统计学的位置：

$$Z = \frac{x_i - \bar{x}}{s} = \frac{75 - 60}{5} = 3$$

即该名学生货币银行学高于全班的均值，且高于 2 个标准差，而统计学高于全班的平均值，且高于 3 个标准差。结论是，该名学生统计学学得较好。

4.4 偏态与峰态的指标度量

4.2 和 4.3 节学习了集中趋势和离散程度的测量，但这两部分的学习仅可以了解数据分布的一些特点，要想全面了解数据分布的特点，还是不够的。例如，给出一组样本数据，可以计算其众数、中位数和平均数 3 个数值，通过 3 个数值的关系，初步了解到数据分布是对称还是非对称，但如果是非对称时，无法知道数据的偏斜程度；如果是对称的，无法知道数据的扁平程度。这时需要学习偏态和峰态的指标度量。

4.4.1 偏态及偏态系数

1. 偏态及偏态系数的定义

定义 4.13 数据分布的不对称，称为偏态。

偏态是对数据分布对称性的测量，如果要测量偏斜程度的话，需要计算偏态系数。

定义 4.14 数据分布不对称的度量值，称为偏态系数，用 SK 表示。

其计算公式为

$$\text{SK} = \frac{1}{n-1} \times \frac{\sum (x_i - \bar{x})^3}{s^3} \tag{4.16}$$

(1) 当偏态系数为 0 时，说明这组数据是对称分布的。

(2) 当偏态系数为正值时，表示这组数据是右偏的，偏态系数越大，偏斜的程度也就越大。

(3) 当偏态系数为负值时，表示这组数据是左偏的，偏态系数越小，偏斜的程度也就越大。

2. Excel 中偏态系数的计算函数

利用 Excel 中的 SKEW 函数可以计算出一组数值数据的偏态系数 SK。其语法为 SKEW(number1,number2,…)。

4.4.2 峰态及峰态系数

1. 峰态及峰态系数的定义

定义 4.15 数据分布的平峰或尖峰程度，称为峰态。

峰态是对数据对称分布的扁平测量，如果要测量扁平程度，需要计算峰态系数。

定义 4.16 数据分布峰态的度量值，称为峰态系数，用 K 表示。

其计算公式为

$$K = \frac{1}{n-1} \times \frac{\sum (x_i - \overline{x})^4}{s^4 - 3} \tag{4.17}$$

峰态通常是相对于标准正态分布而言的。

(1) 当峰态系数为 0 时，说明这组数据是标准正态分布的。

(2) 当峰态系数为正值时，表示这组数据是尖峰的，同时意味着数据比较集中。

(3) 当峰态系数为负值时，表示这组数据是扁平的，同时意味着数据比较分散。

2. Excel 中峰态系数的计算函数

利用 Excel 中的 KURT 函数可以计算出一组数值数据的峰态系数 K。其语法为 KURT(number1,number2,…)。

4.5 案例分析：啤酒市场的调查与分析及 Excel 上机应用——描述性统计指标

本节案例分析主要分析不同性别和不同学历对啤酒的印象分布情况。首先要计算出啤酒综合印象分数，操作过程如下。

打开“自动统计调查结果”工作表，在 Z 列之后插入一列“啤酒印象分数”，同时设非常不同意为 1 分，不同意为 2 分，中立为 3 分，同意为 4 分，非常同意为 5 分，所以对啤酒的印象得分根据“调查问卷”中的第 9 题来计算，计算方法为(1)+(2)+(4)−(3)−(5)，然后拖动 AA2 单元格右下角的填充柄向下复制公式，计算出每位受访者的啤酒综合印象分数，如图 4.3 所示。

AA2 =Q2+S2+W2-U2-Y2

	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA	
1	购买地点	购买地点	饮用量（CC）	饮用量（CC）	增加气氛	增加气氛	解渴	解渴	易发胖	易发胖	营养高	营养高	味苦	味苦	啤酒印象分数	再
2	1	大型超市	4	2000CC以上	2	不同意	3	中立	4	同意	5	非常同意	4	同意	2	
3	2	专卖店	4	2000CC以上	5	非常同意	5	非常同意	3	中立	3	中立	1	非常不同意	9	
4	2	专卖店	3	1000-2000CC	5	非常同意	5	非常同意	4	同意	3	中立	2	不同意	7	
5	2	专卖店	2	500——999CC	4	同意	4	同意	5	非常同意	3	中立	4	同意	2	
6	3	小卖部	3	1000-2000CC	5	非常同意	4	同意	3	中立	3	中立	2	不同意	7	
7	1	大型超市	4	2000CC以上	5	非常同意	5	非常同意	2	不同意	4	同意	1	非常不同意	11	
8	3	小卖部	2	500——999CC	4	同意	2	不同意	5	非常同意	2	不同意	4	同意	-1	
9	2	专卖店	2	500——999CC	5	非常同意	4	同意	3	中立	3	中立	2	不同意	7	
10	2	专卖店	4	2000CC以上	5	非常同意	5	非常同意	2	不同意	4	同意	1	非常不同意	11	
11	2	专卖店	3	1000-2000CC	5	非常同意	4	同意	4	同意	3	中立	2	不同意	6	
12	3	小卖部	2	500——999CC	4	同意	3	中立	4	同意	3	中立	4	同意	2	
13	3	小卖部	3	1000-2000CC	5	非常同意	4	同意	2	不同意	3	中立	1	非常不同意	9	
14	2	专卖店	4	2000CC以上	5	非常同意	5	非常同意	2	不同意	4	同意	1	非常不同意	11	
15	3	小卖部	1	500CC以下	4	同意	2	不同意	5	非常同意	2	不同意	4	同意	-1	
16	1	大型超市	3	1000-2000CC	5	非常同意	4	同意	3	中立	3	中立	2	不同意	7	
17	2	专卖店	4	2000CC以上	5	非常同意	5	非常同意	2	不同意	4	同意	1	非常不同意	11	
18	3	小卖部	4	2000CC以上	5	非常同意	5	非常同意	1	非常不同意	4	同意	1	非常不同意	12	
19	3	小卖部	1	500CC以下	3	中立	3	中立	4	同意	3	中立	3	中立	2	
20	3	小卖部	1	500CC以下	3	中立	3	中立	4	同意	3	中立	4	同意	1	
21	1	大型超市	4	2000CC以上	5	非常同意	5	非常同意	2	不同意	4	同意	1	非常不同意	11	
22	2	专卖店	4	2000CC以上	5	非常同意	5	非常同意	2	不同意	4	同意	1	非常不同意	11	
23	1	大型超市	4	2000CC以上	5	非常同意	5	非常同意	2	不同意	3	中立	1	非常不同意	10	
24	3	小卖部	1	500CC以下	4	同意	3	中立	5	非常同意	3	中立	4	同意	1	
25	1	大型超市	4	2000CC以上	5	非常同意	5	非常同意	2	不同意	3	中立	1	非常不同意	10	
26	3	小卖部	4	2000CC以上	5	非常同意	4	同意	3	中立	3	中立	2	不同意	7	
27	2	专卖店	3	1000-2000CC	4	同意	4	同意	4	同意	3	中立	2	不同意	5	
28	3	小卖部	4	2000CC以上	5	非常同意	5	非常同意	2	不同意	3	中立	1	非常不同意	10	
29	1	大型超市	4	2000CC以上	5	非常同意	2	不同意	4	同意	5	非常同意	1	非常不同意	7	

自动统计调查结果 / 编码设置 / 调查结果数据库 / 样本组成分析 / Sheet1

图 4.3 计算啤酒印象分数

上面已经介绍过，描述一组数据的特性，通常用该组数据的描述性指标来反映其分布的情况，即计算出一组数据的平均数、标准差、众数、中位数等指标，方便用户从“集中程度”和“离散程度”两个角度对样本数据进行观察。本节采用 Excel 软件，使用两种方法来计算。

4.5.1　不同性别的啤酒印象分数分布情况

首先分析一下性别是否会对啤酒印象分数有影响，即计算男、女两组啤酒印象分数描述性指标。这里采用 Excel 统计函数来计算每个指标的数值。

第一步：建立一个新的工作表，命名为“性别对啤酒印象分数的影响分析”，并把性别和啤酒印象分数的数据复制到该工作表中，如图 4.4 所示。

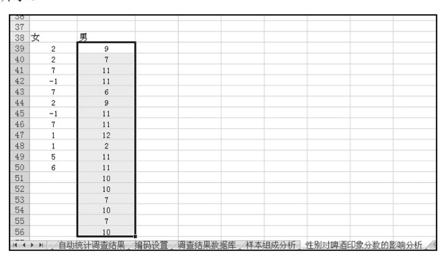


	A	B	C	D	E	F	G	H	I
1	性别	啤酒印象分数							
2	女	2							
3	男	9							
4	男	7							
5	女	2							
6	女	7							
7	男	11							
8	女	-1							
9	女	7							
10	男	11							
11	男	6							
12	女	2							
13	男	9							
14	男	11							
15	女	-1							
16	女	7							
17	男	11							
18	男	12							
19	男	2							
20	女	1							
21	男	11							
22	男	11							
23	男	10							
24	女	1							
25	男	10							
26	男	7							
27	女	5							
28	男	10							
29	男	7							

自动统计调查结果 / 编码设置 / 调查结果数据库 / 样本组成分析 / 性别对啤酒印象分数的影响分析

图 4.4　“性别对啤酒印象分数的影响分析”工作表

第二步：利用 Excel 中的自动筛选功能分别筛选出样本中男、女性各自的啤酒印象分数，如图 4.5 所示。



	A	B
37		
38	女	男
39	2	9
40	2	7
41	7	11
42	-1	11
43	7	6
44	2	9
45	-1	11
46	7	11
47	1	12
48	1	2
49	5	11
50	6	11
51		10
52		10
53		7
54		10
55		7
56		10

自动统计调查结果 / 编码设置 / 调查结果数据库 / 样本组成分析 / 性别对啤酒印象分数的影响分析

图 4.5　筛选结果

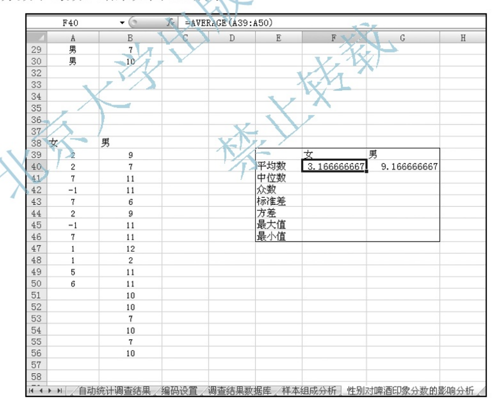
第三步：在“性别对啤酒印象分数的影响分析”工作表中，输入如图 4.6 所示的内容。

	A	B	C	D	E	F	G
37							
38	女	男					
39	2	9				女	男
40	2	7			平均数		
41	7	11			中位数		
42	-1	11			众数		
43	7	6			标准差		
44	2	9			方差		
45	-1	11			最大值		
46	7	11			最小值		
47	1	12					
48	1	2					
49	5	11					
50	6	11					
51		10					
52		10					
53		7					
54		10					
55		7					
56		10					
57							
58							

自动统计调查结果 / 编码设置 / 调查结果数据库 / 样本组成分析 / 性别对啤酒印象分数的影响分析

图 4.6　输入结果

第四步：在 F40 单元格中输入函数“=AVERAGE(A39:A50)”，得到女性样本的啤酒印象分数平均数；同样在 G40 单元格中输入函数“=AVERAGE(B39:B56)”计算男性样本的啤酒印象分数平均数。结果如图 4.7 所示。



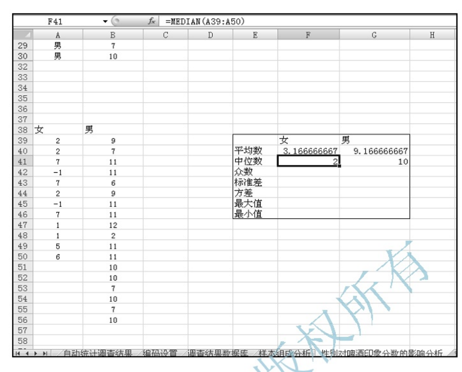
F40　=AVERAGE(A39:A50)

	A	B	C	D	E	F	G	H
29	男	7						
30	男	10						
32								
33								
34								
35								
36								
37								
38	女	男						
39	2	9				女	男	
40	2	7			平均数	3.166666667	9.166666667	
41	7	11			中位数			
42	-1	11			众数			
43	7	6			标准差			
44	2	9			方差			
45	-1	11			最大值			
46	7	11			最小值			
47	1	12						
48	1	2						
49	5	11						
50	6	11						
51		10						
52		10						
53		7						
54		10						
55		7						
56		10						
57								
58								

自动统计调查结果 / 编码设置 / 调查结果数据库 / 样本组成分析 / 性别对啤酒印象分数的影响分析

图 4.7　计算样本的啤酒印象分数平均数

第五步：在 F41 单元格中输入函数“=MEDIAN(A39:A50)”，得到女性样本的啤酒印象分数中位数；同样在 G41 单元格中输入函数“=MEDIAN(B39:B56)”计算男性样本的啤酒印象分数中位数。结果如图 4.8 所示。



F41 =MEDIAN(A39:A50)

	A	B	C	D	E	F	G	H
29	男	7						
30	男	10						
32								
33								
34								
35								
36								
37								
38	女	男						
39	2	9				女	男	
40	2	7			平均数	3.166666667	9.166666667	
41	7	11			中位数	2	10	
42	-1	11			众数			
43	7	6			标准差			
44	2	9			方差			
45	-1	11			最大值			
46	7	11			最小值			
47	1	12						
48	1	2						
49	5	11						
50	6	11						
51		10						
52		10						
53		7						
54		10						
55		7						
56		10						
57								
58								

自动统计调查结果 / 编码设置 / 调查结果数据库 / 样本组成分析 / 性别对啤酒印象分数的影响分析

图 4.8 计算样本的啤酒印象分数中位数

第六步：在 F42 单元格中输入函数“=MODE(A39:A50)”，得到女性样本的啤酒印象分数众数；同样在 G42 单元格中输入函数“=MODE(B39:B56)”，计算男性样本的啤酒印象分数众数。结果如图 4.9 所示。



F42 =MODE(A39:A50)

	A	B	C	D	E	F	G	H
29	男	7						
30	男	10						
32								
33								
34								
35								
36								
37								
38	女	男						
39	2	9				女	男	
40	2	7			平均数	3.166666667	9.166666667	
41	7	11			中位数	2	10	
42	-1	11			众数	2	11	
43	7	6			标准差			
44	2	9			方差			
45	-1	11			最大值			
46	7	11			最小值			
47	1	12						
48	1	2						
49	5	11						
50	6	11						
51		10						
52		10						
53		7						
54		10						
55		7						
56		10						
57								
58								

自动统计调查结果 / 编码设置 / 调查结果数据库 / 样本组成分析 / 性别对啤酒印象分数的影响分析

图 4.9 计算样本的啤酒印象分数众数

第七步：在 F43 单元格中输入函数“=STDEV(A39:A50)”，得到女性样本的啤酒印象分数标准差；同样在 G43 单元格中输入函数“=STDEV(B39:B56)”，计算男性样本的啤酒印象分数标准差。结果如图 4.10 所示。

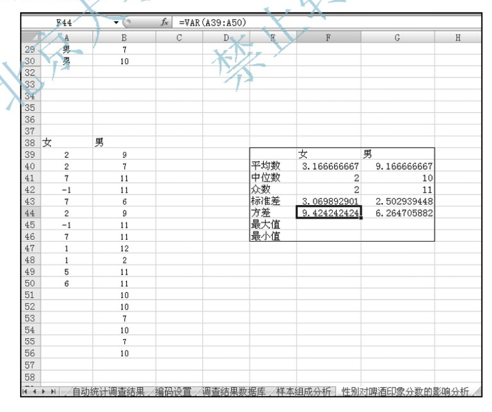
F43 =STDEV(A39:A50)

	A	B	C	D	E	F	G	H
29	男	7						
30	男	10						
32								
33								
34								
35								
36								
37								
38	女	男						
39	2	9				女	男	
40	2	7			平均数	3.166666667	9.166666667	
41	7	11			中位数	2	10	
42	-1	11			众数	2	11	
43	7	6			标准差	3.069892901	2.502939448	
44	2	9			方差			
45	-1	11			最大值			
46	7	11			最小值			
47	1	12						
48	1	2						
49	5	11						
50	6	11						
51		10						
52		10						
53		7						
54		10						
55		7						
56		10						
57								
58								

自动统计调查结果 / 编码设置 / 调查结果数据库 / 样本组成分析 / 性别对啤酒印象分数的影响分析

图 4.10 计算样本的啤酒印象分数标准差

第八步：在 F44 单元格中输入函数“=VAR(A39:A50)”，得到女性样本的啤酒印象分数方差；同样在 G44 单元格中输入函数“=VAR(B39:B56)”，计算男性样本的啤酒印象分数方差。结果如图 4.11 所示。



F44 =VAR(A39:A50)

	A	B	C	D	E	F	G	H
29	男	7						
30	男	10						
32								
33								
34								
35								
36								
37								
38	女	男						
39	2	9				女	男	
40	2	7			平均数	3.166666667	9.166666667	
41	7	11			中位数	2	10	
42	-1	11			众数	2	11	
43	7	6			标准差	3.069892901	2.502939448	
44	2	9			方差	9.424242424	6.264705882	
45	-1	11			最大值			
46	7	11			最小值			
47	1	12						
48	1	2						
49	5	11						
50	6	11						
51		10						
52		10						
53		7						
54		10						
55		7						
56		10						
57								
58								

自动统计调查结果 / 编码设置 / 调查结果数据库 / 样本组成分析 / 性别对啤酒印象分数的影响分析

图 4.11 计算样本的啤酒印象分数方差

第九步：在 F45 单元格中输入函数“=MAX(A39:A50)”，得到女性样本的啤酒印象分数最大值；同样在 G45 单元格中输入函数“=MAX(B39:B56)”，计算男性样本的啤酒印象分数最大值。结果如图 4.12 所示。

F45 =MAX(A39:A50)

	A	B	C	D	E	F	G	H
29	男	7						
30	男	10						
32								
33								
34								
35								
36								
37								
38	女	男						
39	2	9				女	男	
40	2	7			平均数	3.166666667	9.166666667	
41	7	11			中位数	2	10	
42	-1	11			众数	2	11	
43	7	6			标准差	3.069892901	2.502939448	
44	2	9			方差	9.424242424	6.264705882	
45	-1	11			最大值	7	12	
46	7	11			最小值			
47	1	12						
48	1	2						
49	5	11						
50	6	11						
51		10						
52		10						
53		7						
54		10						
55		7						
56		10						
57								
58								

自动统计调查结果 / 编码设置 / 调查结果数据库 / 样本组成分析 / 性别对啤酒印象分数的影响分析

图 4.12　计算样本的啤酒印象分数最大值

第十步：在 F46 单元格中输入函数“=MIN(A39:A50)”，得到女性样本的啤酒印象分数最小值；同样在 G46 单元格中输入函数“=MIN(B39:B56)”，计算男性样本的啤酒印象分数最小值。结果如图 4.13 所示。

F46 =MIN(A39:A50)

	A	B	C	D	E	F	G	H
29	男	7						
30	男	10						
32								
33								
34								
35								
36								
37								
38	女	男						
39	2	9				女	男	
40	2	7			平均数	3.166666667	9.166666667	
41	7	11			中位数	2	10	
42	-1	11			众数	2	11	
43	7	6			标准差	3.069892901	2.502939448	
44	2	9			方差	9.424242424	6.264705882	
45	-1	11			最大值	7	12	
46	7	11			最小值	-1	2	
47	1	12						
48	1	2						
49	5	11						
50	6	11						
51		10						
52		10						
53		7						
54		10						
55		7						
56		10						
57								
58								

自动统计调查结果 / 编码设置 / 调查结果数据库 / 样本组成分析 / 性别对啤酒印象分数的影响分析

图 4.13　计算样本的啤酒印象分数最小值

第十一步：最后得到如图 4.14 所示的结果。

	女	男
平均数	3.166666667	9.166666667
中位数	2	10
众数	2	11
标准差	3.069892901	2.502939448
方差	9.424242424	6.264705882
最大值	7	12
最小值	-1	2

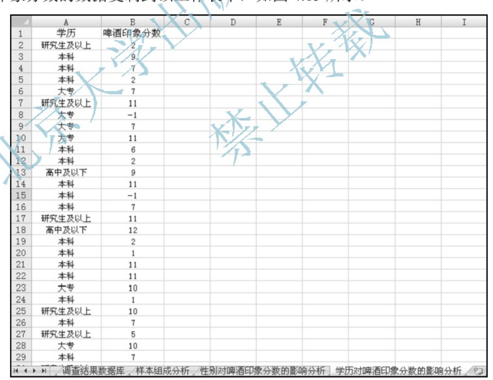
图 4.14　统计结果报表

根据图 4.14 的描述性统计结果报表可知，男性对啤酒的平均印象分数远远高于女性(9.17＞3.17)，即男性对啤酒的印象较佳，而女性对啤酒的印象较差。但虽然从描述统计分析结论来看是这样，但不能准确说明对啤酒印象与性别之间有关，必须还要借助于另外的分析工具进行分析检验。

4.5.2　不同学历的啤酒的印象分数分布情况

分析不同学历的啤酒印象分数的分布情况，这里采用数据中的描述统计分析，不再用统计函数来计算各指标数值。操作过程如下。

第一步：建立一个新的工作表，命名为“学历对啤酒印象分数的影响分析”，然后把学历和啤酒印象分数的数据复制到该工作表中，如图 4.15 所示。



	A	B
1	学历	啤酒印象分数
2	研究生及以上	2
3	本科	9
4	本科	7
5	本科	2
6	大专	7
7	研究生及以上	11
8	大专	-1
9	大专	7
10	大专	11
11	本科	6
12	本科	2
13	高中及以下	9
14	本科	11
15	本科	-1
16	本科	7
17	研究生及以上	11
18	高中及以下	12
19	本科	2
20	本科	1
21	本科	11
22	本科	11
23	大专	10
24	本科	1
25	研究生及以上	10
26	本科	7
27	研究生及以上	5
28	大专	10
29	本科	7

调查结果数据库 / 样本组成分析 / 性别对啤酒印象分数的影响分析 / 学历对啤酒印象分数的影响分析

图 4.15　“学历对啤酒印象分数的影响分析”工作表

第二步：利用 Excel 中的自动筛选功能分别筛选出样本中不同学历的各自啤酒印象分数，如图 4.16 所示。

第三步：单击“数据”→“分析”→“数据分析”按钮，弹出“数据分析”对话框，在“分析工具”列表中选择“描述统计”选项，如图 4.17 所示。

40	高中及以下	大专	本科	研究生及以上
41	9	7	9	2
42	12	-1	7	11
43		7	2	11
44		11	6	10
45		10	2	5
46		10	11	10
47			-1	
48			7	
49			2	
50			1	
51			11	
52			11	
53			1	

调查结果数据库 / 样本组成分析 / 性别对啤酒印象分数的影响分析 / 学历对啤酒印象分数的影响分析

图 4.16　样本中不同学历的各自啤酒印象分数

第四步：单击“确定”按钮后，弹出“描述统计”对话框，需要填好数据的输入区域 A40:D56，填好输出区域 F47，选择统计指标，即选择汇总统计，如图 4.18 所示。

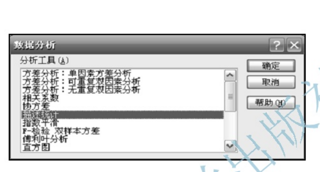


图 4.17　“数据分析”对话框

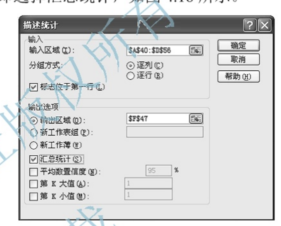


图 4.18　“描述统计”对话框

第五步：单击“确定”按钮，统计结果如图 4.19 所示。

高中及以下		大专		本科		研究生及以上	
平均	10.5	平均	7.3333333	平均	5.5625	平均	8.1666667
标准误差	1.5	标准误差	1.8012341	标准误差	0.98305964	标准误差	1.5365907
中位数	10.5	中位数	8.5	中位数	6.5	中位数	10
众数	#N/A	众数	7	众数	7	众数	11
标准差	2.1213203	标准差	4.4121046	标准差	3.93223855	标准差	3.7638633
方差	4.5	方差	19.466667	方差	15.4625	方差	14.166667
峰度	#DIV/0!	峰度	3.1789501	峰度	-1.2151768	峰度	-0.327031
偏度	#DIV/0!	偏度	-1.716945	偏度	-0.0917788	偏度	-1.172762
区域	3	区域	12	区域	12	区域	9
最小值	9	最小值	-1	最小值	-1	最小值	2
最大值	12	最大值	11	最大值	11	最大值	11
求和	21	求和	44	求和	89	求和	49
观测数	2	观测数	6	观测数	16	观测数	6

图 4.19　统计结果

从图 4.19 的统计结果可以看出，高中及以下对啤酒的印象分数最高，其次是研究生及以上，再次是大专学历，最后是本科学历。高中及以下这组由于数据过少，因此出现了众数、峰度、偏度无值或错误标志。从这 30 个观测值可以看出，学历对啤酒印象分数有显著的影响，但由于样本的随机性，因此还不能非常肯定地说学历对啤酒印象分数有影响，还需要借助其他的检验方法来肯定。

习　　题

一、填空题

1．一组数据中出现频数最多的变量值为(　　)。

2．一组数据排序后处于中间位置上所对应的变量值为(　　)。

3．上四分位数减下四分位数的结果，称为(　　)。

4．当一组数据的众数、中位数和平均数相等时，这组数据的分布(　　)。

5．一组数据的离散系数为 0.4 时，该组数据中每个变量值变为原变量值的 2 倍，此时数据的离散系数为(　　)。

6．测量数据是否是对称分布的统计量是(　　)。

7．一组数据的方差为 16，平均数为 2，则这组数据的离散系数为(　　)。

8．峰态通常是与标准正态分布相比较而言的，如果一组数据是服从标准正态分布，则峰态系数为(　　)。

9．如果一组数据比较集中，则这组数据的峰态系数为(　　)。

10．标准分数的公式为(　　)。

二、单项选择题

1．一组数据排序后，处于 25%位置上所对应的变量值为(　　)。

A．众数　　B．上四分位数　　C．下四分位数　　D．中位数

2．一组数据排序后，处于 75%位置上所对应的变量值为(　　)。

A．众数　　B．上四分位数　　C．下四分位数　　D．中位数

3．8 个数据的平均数是 10，其中一个数为 6，那么其余 7 个数的平均数是(　　)。

A．10.6　　B．10.2　　C．10.7　　D．10.9

4．有两组样本，两组样本平均数相等，其中第一组的样本方差为 $s_1^2=6$，$s_2^2=8$，试问两组样本的波动情况是(　　)。

A．两组波动情况相同　　B．第二组比第一组波动程度大

C．第一组比第二组波动程度大　　D．无法比较

5．当一组数据是左偏分布时，这组数据的众数、中位数和平均数的关系是(　　)。

A．$M_0=M_e=\bar{x}$　　B．$\bar{x}<M_e<M_0$

C．$M_0<M_e<\bar{x}$　　D．$M_e<M_0<\bar{x}$

6．当一组数据是右偏分布时，这组数据的众数、中位数和平均数的关系是(　　)。

A．$M_0=M_e=\bar{x}$　　B．$\bar{x}<M_e<M_0$

C．$M_0<M_e<\bar{x}$　　D．$M_e<M_0<\bar{x}$

7．当一组数据是对称分布时，这组数据的众数、中位数和平均数的关系是(　　)。

A．$M_0=M_e=\bar{x}$　　B．$\bar{x}<M_e<M_0$

C．$M_0<M_e<\bar{x}$　　D．$M_e<M_0<\bar{x}$

8．中位数主要适用于测量顺序数据的集中趋势，但可以测量(　　)。

A．分类数据　　B．数值数据

C．分类数据和数值数据　　D．以上都不能

9. 如果一个数据的标准分数为-4，表明数据(　　)。
   A. 比平均数高出 4 个标准差
   B. 比平均数低出 4 个标准差
   C. 比平均数高出 4 个方差
   D. 比平均数低出 4 个方差
10. 比较两组数据的离散程度采用的统计量是(　　)。
   A. 方差　　B. 标准差　　C. 四分位数　　D. 离散系数
11. 下列统计量不受极端值的影响的是(　　)。
   A. 众数　　B. 平均数　　C. 加权平均数　　D. 方差
12. 比较两组数据离散程度使用离散系数，其原因是(　　)。
   A. 两组数据的个数不同
   B. 两组数据的平均数不同
   C. 两组数据的方差不同
   D. 两组数据的数据水平不同或计量单位不同
13. 两组数据的均值不等，但方差相等，则离散程度(　　)。
   A. 均值大的，离散程度小　　B. 均值小的，离散程度大
   C. 两组数据的离散程度相同　　D. 无法比较
14. 下列指标受极端值的影响的是(　　)。
   A. 中位数　　B. 众数　　C. 平均数　　D. 四分位数
15. 如果一个数据的标准分数为 4，表明数据比平均数(　　)。
   A. 高出 4 个标准差　　B. 低出 4 个标准差
   C. 高出 4 个方差　　D. 低出 4 个方差

三、多项选择题

1. 下列关于众数的叙述，正确的有(　　)。
   A. 一组数据可能存在多个众数
   B. 众数主要适用于分类数据的集中趋势度量，也可以测量顺序数据和数值数据
   C. 众数不受极端值的影响
   D. 一组数据存在唯一的众数
2. 下列有关离散系数的叙述，正确的有(　　)。
   A. 离散系数主要是比较多组数据的离散程度
   B. 离散系数可以同时消除数据的水平和计量单位对标准差的影响
   C. 离散系数大的离散程度大，离散系数小的离散程度小
   D. 离散系数大的离散程度小，离散系数小的离散程度大
3. 如果偏态系数为正值，则表明数据的分布是(　　)。
   A. 右偏的　　B. 非对称的　　C. 左偏的　　D. 对称的
4. 异众比率是衡量一组数据的离散程度，它可以测量(　　)。
   A. 分类数据　　B. 顺序数据　　C. 数值数据　　D. 以上都不可以
5. 分组数据计算平均数时，使用加权平数，加权平均数受(　　)影响。
   A. 频数　　B. 组中值　　C. 最大值　　D. 最小值

6．测量数值数据的集中趋势的统计量有(　　)。

A．众数　　B．中位数　　C．平均数　　D．以上都不对

7．某小区准备对其服务采取新的收费标准，为此，它随机抽取了该小区 100 户居民进行调查，其中表示赞成的有 23 户，中立的有 20 户，不赞成的有 57 户，试问描述该组数据的集中趋势统计量有(　　)。

A．众数　　B．中位数　　C．平均数　　D．方差

8．下列指标不受极端值的影响的是(　　)。

A．众数　　B．中位数　　C．平均数　　D．方差

四、名词解释

1．众数和中位数。

2．简单平均数和加权平均数。

3．方差和标准差。

4．离散系数。

5．偏度和峰度。

五、计算题

1．某班共有 25 名学生，期末经济学课程的成绩分数分别为 68、73、66、78、86、74、60、89、64、90、69、67、76、62、81、63、68、81、81、81、81、70、60、87、64。

试回答以下问题：

(1) 计算该组数据的众数。

(2) 计算该组数据的中位数及四分位数。

(3) 计算该组数据的平均数和方差。

(4) 写出以上 3 个问题 Excel 的计算过程。

2．某班期末共进行经济学和统计学两门课程的考试，全班经济学成绩的平均分数为 80 分，标准差是 15 分；全班统计学成绩的平均分数为 65 分，标准差为 5 分。一名学生在经济学成绩为 85 分，统计学成绩为 70 分，试问，该名学生在哪门课程的考试表现理想？

3．一项关于 4 岁儿童身高状况的研究发现，女童的平均身高为 105cm，标准差为 5cm；男童的平均身高为 107cm，标准差为 10cm，试回答以下问题：

(1) 女童和男童身高差异哪个大？为什么？

(2) 如果该题的单位由“cm”转为“m”，女童和男童身高差异哪个大？为什么？

4．某企业生产日光灯，随机抽取了 120 个日光灯，测得寿命数据如表 4-7 所示，试计算该分组数据的平均数和标准差。

表 4-7　日光灯寿命测试数据

按寿命分组	频数
500～1 000	19
1 000～1 500	30
1 500～2 000	42
2 000～2 500	18
2 500 以上	11
合计	120

5. 一种产品的组装方法有两种，现要从两种方法中选出一种单位时间组装产品最多的方法，随机抽取 10 个工人，并让他们分别用这两种方法进行产品组装，单位时间组装产品个数数据如表 4-8 所示。试问，采用什么指标比较两种组装方法的离散程度？如果是你选择，你会选择哪种组装方法？

表 4-8　两种方法单位时间内组装产品个数

方法 1	方法 2
164	129
167	125
170	126
165	130
168	128
162	127
160	130
168	128
171	127
165	131

第5章 参数估计

教学目标

1. 掌握几个重要的统计分布。
2. 了解参数估计的基本理论。
3. 掌握一个总体参数的区间估计。
4. 掌握样本容量的确定。

引入案例

全校在校大学生每月平均消费支出

为了解全国在校大学生每月平均消费支出，光华学院经济学专业的4名本科生对全校部分本科生做了问卷调查。调查的对象为光华学院在校本科生，调查的内容包括每月平均消费支出、支出的途径、支出结构等，调查问卷由调查员直接到宿舍发放并当场收回。对4个年级中每个年级各发放了60份，其中男女生各占一半。共收回有效问卷200份。其中有关月平均消费支出方面的数据整理见表5-1所示。

表5-1　月平均消费支出调查数据整理表

回答类别	人数/人	频率/%
500元以下	32	16
600～800元	80	40
800～1 000元	55	27.5
1 000元以上	33	16.5
合计	200	100

根据表5-1数据计算的平均月消费支出为 $\overline{x} = 749$ 元，试问全校学生每月平均消费支出是多少？作出估计的理论依据是什么？本章的内容就将回答这些问题。

5.1　几个重要的统计分布

在学习推断统计的两个方法，即参数估计和假设检验前，必须要学习统计学中的几个重要的分布，因为在参数估计和假设检验时要用到这几个重要的分布，由于这几个重要的分布，在《概率与数理统计》中已经学习过，这里只是简单地介绍一下。

5.1.1 正态分布

定义 5.1　设随机变量 X，如果其概率密度函数为

$$f(x)=\frac{1}{\sqrt{2\pi\sigma^2}}\mathrm{e}^{-\frac{(x-\mu)^2}{2\sigma^2}} \qquad (-\infty < x < \infty)$$

式中，μ 为随机变量 X 的均值，它可以为任意的实数；σ^2 为随机变量 X 的方差。则称随机变量 X 服从正态分布。即表示为

$$X \sim N(\mu,\sigma^2) \tag{5.1}$$

正态分布是关于 $x=\mu$ 对称分布；正态分布中 σ^2 代表分布的曲线扁平还是陡峭的，当其值越小时，曲线越陡峭，其值越大时，曲线越扁平。正态分布的随机变量的线性组合后的随机变量也服从正态分布。

5.1.2 标准正态分布

定义 5.2　标准正态分布是正态分布的特例，当定义 5.1 中，有 $\mu=0,\sigma^2=1$，则称随机变量服从标准正态分布，即

$$X \sim N(0,1) \tag{5.2}$$

此时，标准正态分布是关于 y 轴对称分布的图形。

由于标准正态分布是正态分布的特例，因此可以将任何一个服从一般正态分布的随机变量 $X \sim N(\mu,\sigma^2)$ 转换成标准正态分布 $N(0,1)$，转换公式为

$$Z=\frac{X-\mu}{\sigma} \tag{5.3}$$

转换后的 Z 是一个服从标准正态分布的随机变量，即 $Z \sim N(0,1)$。

5.1.3 χ^2(卡方)分布

定义 5.3　设一组相互独立的随机变量 $X_1,X_2,\cdots,X_n$，且随机变量 $X_i \sim N(0,1)$，则随机变量：

$$\chi^2=\sum X_i^2 \sim \chi^2(n) \tag{5.4}$$

式中，n 为自由度。

χ^2 是由标准正态分布的平方加和得到的随机变量，所以随机变量 χ^2 为非负数，即 $\chi^2 \geqslant 0$；同时 χ^2 是非对称分布。

5.1.4 t 分布

定义 5.4　设有一随机变量 X 服从标准正态分布，另一随机变量 Y 服从 χ^2 分布，即 $X \sim N(0,1)$，$Y \sim \chi^2(n)$，则有

$$t=\frac{X}{\sqrt{Y/n}} \sim t(n) \tag{5.5}$$

式中，n 为 t 的自由度。

由定义可知，t 分布是一个关于 y 轴对称的分布图形。

5.1.5 F分布

定义 5.5 设有一随机变量 X 服从 χ^2 分布，另一随机变量 Y 服从 χ^2 分布，即 $X \sim \chi^2(n_1)$，$Y \sim \chi^2(n_2)$，则有

$$F = \frac{X / n_1}{Y / n_2} \sim F(n_1, n_2) \tag{5.6}$$

式中，n_1 为第一自由度；n_2 为第二自由度。

由定义同样可以得出随机变量 F 是非负数，即 $F \geqslant 0$，非对称分布。

【例 5.1】 随机变量 $t \sim t(n)$，证明 $t^2 \sim F(1,n)$。

证明：根据 $t \sim t(n)$，由 t 分布定义可知，由一个标准正态分布和一个 χ^2 构造的，所以设有一随机变量 X 服从标准正态分布，即 $X \sim N(0,1)$；另一随机变量 Y 服从 χ^2 分布，即 $Y \sim \chi^2(n)$，则有

$$t = \frac{X}{\sqrt{Y / n}} \sim t(n)$$

则有

$$t^2 = \frac{X^2}{Y / n}$$

其中 $Y \sim \chi^2(n)$，再看 X^2，因为 $X \sim N(0,1)$，那么 $X^2 \sim \chi^2(1)$，所以有

$$t^2 \sim F(1,n)$$

5.2 样本抽样分布

前面已经介绍过，统计学主要是分析数据，得出数据的规律性，即得出研究对象(总体)的特征(参数)，总体的参数有 μ, π, σ^2。一般情况，想得出总体的这些参数，需要收集总体的数据，而总体的数据是不易收集，甚至是收集不到的，所以只能利用推断统计，先来计算样本的统计量值 $\bar{x}, p, s^2$，推断出总体的参数 μ, π, σ^2。那么必须要学习这种推断统计的理论依据，即样本的抽样分布。

定义 5.6 重复选取样本量为 n 的样本，由该统计量的所有可能取值形成的概率分布，称为样本抽样分布。

样本抽样分布指的是样本的统计量分布，本书主要介绍样本均值 $\bar{x}$ 的分布，样本比例 p 的分布和样本方差 s^2 的分布，因为统计学最为关心是总体的均值、比例和方差，而 3 个参数往往是利用推断统计，从样本的均值、比例和方差进行估计的。

这里需要注意的是，在实务中，抽样采取的是重复抽样，所以下面的研究都是在重复抽样的基础上进行的。

5.2.1 样本均值的抽样分布

定义 5.7 重复选取样本量为 n 的样本，由样本均值的所有可能取值形成的概率分布，称为样本均值的抽样分布。

下面以一个简单的例子来推导样本均值的抽样分布。

【例 5.2】 设一个总体含有 4 个元素(个体)，即总体元素个数 $N=4$，4 个元素的取值分别为 $x_1=2$ 、 $x_2=3$ 、 $x_3=4$ 、 $x_4=5$ 。从总体中采取重复抽样方法抽取样本量为 $n=2$ 的随机样本，写出样本均值 $\bar{x}$ 的抽样分布。

解：从总体分布情况看，总体的分布为均匀分布，即 x_i 取每一个值的概率都相同。计算总体均值和方差分别为

总体均值：

$$\mu=\frac{\sum_{i=1}^{4}x_i}{N}=\frac{14}{4}=3.5$$

总体方差：

$$\sigma^2=\frac{\sum_{i=1}^{4}(x_i-\mu)^2}{4}=\frac{5}{4}=1.25$$

从总体中采取重复抽样方法抽取样本量为 $n=2$ 的随机样本，共有 16 个可能的样本，见表 5-2 所示。

表 5-2 样本的所有情况

样本	样本中的元素	样本均值	概率
1	2,2	2.0	1/16
2	2,3	2.5	1/16
3	2,4	3	1/16
4	2,5	3.5	1/16
5	3,2	2.5	1/16
6	3,3	3.0	1/16
7	3,4	3.5	1/16
8	3,5	4.0	1/16
9	4,2	3.0	1/16
10	4,3	3.5	1/16
11	4,4	4.0	1/16
12	4,5	4.5	1/16
13	5,2	3.5	1/16
14	5,3	4.0	1/16
15	5,4	4.5	1/16
16	5,5	5.0	1/16

从表 5-2 中可以得到，样本均值 $\bar{x}$ 的取值有以下几种情况，且相应的概率分布见表 5-3 所示。

表 5-3 样本均值的概率分布

样本均值 $\bar{x}$	概率
2.0	1/16
2.5	2/16
3.0	3/16
3.5	4/16
4.0	3/16
4.5	2/16
5.0	1/16

将样本均值 $\bar{x}$ 的分布绘制成图，如图 5.1 所示，发现样本均值 $\bar{x}$ 是对称分布。

由图 5.1 可知，样本均值 $\bar{x}$ 是关于均值 3.5 对称的，而总体的均值也是 3.5，所以说样本均值 $\bar{x}$ 的均值与总体的均值相等。计算得出样本的方差为 0.625，是总体方差的一半，即 1/2，而 2 又是样本容量，所以得出 $D(\bar{x})=\frac{1}{n}\sigma^2$，即

$$\bar{X}\sim N(\mu,\frac{1}{n}\sigma^2) \tag{5.7}$$

样本的均值抽样分布与原有总体的分布和样本容量 n 大小是有关的。

(1) 当总体是正态分布时，无论样本量的大小，样本均值 $\bar{x}$ 都服从正态分布。

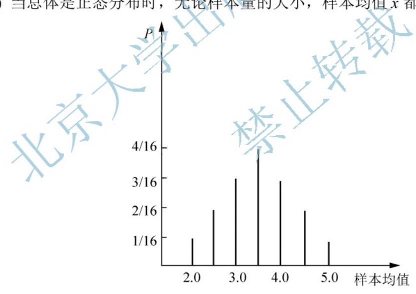


图 5.1 样本均值分布图

(2) 当总体是非正态分布时，样本量为大样本时，样本均值 $\bar{x}$ 也服从正态分布。

(3) 当总体是非正态分布时，样本量为小样本时，样本均值 $\bar{x}$ 不服从正态分布。

本书只考虑前两种情况，即样本均值 $\bar{x}$ 都服从正态分布的情况。

5.2.2 样本比例的抽样分布

在经济管理中，经常要使用到比例，如想估计一批产品的次品率，往往要从样本的比例 p 推断总体的比例 π 。

定义 5.8 总体(或样本)中具有某种属性的单位数与全部单位数的比值，称为比例。

定义 5.9　重复选取样本量为 n 的样本，由样本比例所形成的所有可能取值概率分布，称为样本比例的抽样分布。

样本比例的抽样分布与样本均值的研究方法相似，本书只考虑大样本的情况下，最后推导出样本比例 p 的抽样分布为

$$p \sim N(\pi, \frac{\pi(1-\pi)}{n}) \tag{5.8}$$

5.2.3　样本方差的抽样分布

定义 5.10　重复选取样本量为 n 的样本，由样本方差所有可能取值形成的概率分布，称为样本方差的抽样分布。

$$\chi^2 = \frac{(n-1)s^2}{\sigma^2} \sim \chi^2(n-1) \tag{5.9}$$

证明：

$$\chi^2 = \frac{(n-1)\dfrac{\sum(x_i-\overline{x})^2}{n-1}}{\sigma^2} = \frac{\sum(x_i-\overline{x})^2}{\sigma^2} = \sum(\frac{x_i-\overline{x}}{\sigma})^2 \sim \chi^2(n-1)$$

所以有

$$\chi^2 = \frac{(n-1)s^2}{\sigma^2} \sim \chi^2(n-1)$$

5.3　参数估计的基本理论

5.3.1　参数估计的含义

定义 5.11　根据样本数据提供的信息来推断总体的参数，称为参数估计。

前面已经介绍过参数是描述总体特征的，而最关心总体特征的有总体均值，总体方差和总体比例，如一批灯泡的平均寿命(总体均值)、投资组合的风险(总体方差)和一批灯泡的次品率(总体比例)。这些总体的特征往往是利用样本的数据推断出来，即通常用样本均值 $\overline{x}$ 估计总体均值 μ；用样本方差 s^2 估计总体方差 σ^2；用样本比例 p 估计总体比例 π 等。

5.3.2　参数估计的几个基本概念

1. 估计量与估计值

定义 5.12　估计总体参数 θ 的估计量的名称，称为估计量，用符号 $\hat{\theta}$ 表示。

例如：样本均值 $\overline{x}$、样本比例 p、样本方差 s^2 等都是一个估计量。

定义 5.13　估计总体参数时计算出来的估计量的具体数值，称为估计值。

例如：要估计某学院学生考试的平均成绩，这时该学院是研究的总体，其平均数值 μ 为参数。随机在该学院抽取了一个(班级)样本，该班级的平均数 $\overline{x}$，根据这个样本平均分数估计整个学院的平均分数，所以 $\overline{x}$ 就是一个估计量。假定计算得出样本平均分数为 80 分，那么这个 80 分就是估计量的具体值，称为估计值。

2. 点估计与区间估计

参数估计的方法有点估计和区间估计

1) 点估计

(1) 点估计的定义。

定义 5.14　用样本统计量 $\hat{\theta}$ 的某个取值直接作为总体参数 θ 的估计值，称为参数估计的点估计。

例如，用样本均值 $\bar{x}$ 直接作为总体均值 μ 的估计值，用样本比例 p 直接作为总体比例 π 的估计值。用样本方差 s^2 直接作为总体方差 σ^2 的估计值。

(2) 点估计的优缺点。

① 点估计的优点。点估计的计算较简单，易于理解。例如，假定要估计一批灯泡产品的合格率，从这批灯泡中随机抽取 20 只灯泡，如果抽样结果合格率为 96%，那么将 96%直接作为这批产品的合格率的估计值。

② 点估计的缺点。虽然在重复抽样的条件下，点估计的均值可能等于总体真值，但由于样本是随机的，因此抽出一个具体的样本所得到的估计值很可能不同于总体真值，即表明一个具体的点估计值无法给出点估计的可靠性，因此不能完全依赖于一个点估计值，而是应该围绕点估计值构造出总体参数的一个区间，即区间估计。

2) 区间估计

定义 5.15　在点估计值的基础上，给出总体参数估计的一个范围，称为参数的区间估计。

例如，一名高考学生在考完英语后，估计自己的成绩，估计的成果为 90%的概率，成绩为 120～130 分。这个估计就是区间估计。其概率的表达形式为 $p(120 \leqslant X \leqslant 130)=90\%$，即该名学生的英语成绩是未知参数，这个未知参数有 90%的概率为 120～130。同时也表明让人以某种程度上确信这个区间会包真正的总体参数，所以给它取名为置信区间。

定义 5.16　由样本统计量所构造的总体参数的估计区间，称为置信区间，其中区间的最小值称为置信下限，最大值称为置信上限。

其中置信区间是在以概率为 90%的水平存在，这里的概率取名为置信水平。

定义 5.17　如果将构造置信区间的步骤重复多次，那么区间中包含总体参数真值的次数所占的比率，称为置信水平。

在构造置信区间时，可以用所希望的任意值作为置信水平。但通常情况下，置信水平取 90%、95%和 99%。

3. 标准误差

定义 5.18　样本统计量的抽样分布的标准差，称为统计量的标准误差。标准误差是衡量统计量的离散程度的。

5.3.3 评价估计量的标准

在参数估计中，一般是用样本估计量$\hat{\theta}$作为总体参数θ的估计。实际上，用于估计θ的估计量有很多，如可以用样本均值作为总体均值的估计量，也可以用样本中位数作为总体均值的估计量等。那么，究竟用样本的哪种估计量作为总体参数的估计呢？自然要用估计效果最好的哪种估计量。什么样的估计量才算是一个好的估计量呢？这就需要有一定的评价标准。评价估计量的标准，主要有以下几个。

1. 无偏性

定义 5.19 估计量抽样分布的数学期望等于被估计的总体参数，称为无偏性。即设总体参数为θ，所选择的估计量为$\hat{\theta}$，如果$E(\hat{\theta})=\theta$，则称$\hat{\theta}$为θ的无偏估计量。

通常用样本均值$\bar{x}$估计总体均值μ；用样本方差s^2估计总体方差σ^2；用样本比例p估计总体比例π等，即样本均值$\bar{x}$是总体均值μ的无偏估计；样本方差s^2是总体方差σ^2的无偏估计；样本比例p是总体比例π的无偏估计。在讨论抽样分布时，曾经提到$E(\bar{x})=\mu$和$E(p)=\pi$，同样可以证明$E(s^2)=\sigma^2$。

证明：$E(s^2)=\sigma^2$。

$$
\begin{aligned}
E(s^2) &= E\left(\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1}\right) \\
&= \frac{1}{n-1}E\sum_{i=1}^{n}(x_i^2-2\bar{x}x_i+\bar{x}^2) \\
&= \frac{1}{n-1}E\left(\sum_{i=1}^{n}x_i^2-2\bar{x}\sum_{i=1}^{n}x_i+n\bar{x}^2\right) \\
&= \frac{1}{n-1}E\left(\sum_{i=1}^{n}x_i^2-2\bar{x}n\bar{x}+n\bar{x}^2\right) \\
&= \frac{1}{n-1}E\left(\sum_{i=1}^{n}x_i^2-n\bar{x}^2\right)=\sigma^2
\end{aligned}
$$

注意：一个参数的无偏估计量并不是唯一的，如$\bar{x}=\dfrac{x_1+x_2+\cdots+x_n}{n}$和$\bar{x}'=\dfrac{a_1x_1+a_2x_2+\cdots+a_nx_n}{a_1+a_2+\cdots+a_n}$（其中$\sum a_i\neq 0$）都是总体均值$\mu$的无偏估计量。前面已经得出$E(\bar{x})=\mu$，则$E(\bar{x}')=\mu$证明如下：

$$
\begin{aligned}
E(\bar{x}') &= E\left(\frac{a_1x_1+a_2x_2+\cdots+a_nx_n}{a_1+a_2+\cdots+a_n}\right)=\frac{\sum E(a_ix_i)}{\sum a_i} \\
&= \frac{\sum a_iE(x_i)}{\sum a_i}=\frac{\sum a_i\mu}{\sum a_i}=\mu
\end{aligned}
$$

2. 有效性

由于一个参数的无偏估计并不是唯一的，要想得出参数的最好估计量，就要看估计量与参数的离散程度，即一个无偏的估计量并不意味着它就非常接近被估计的参数，它还必须与总体参数的离散程度比较小。也就是说，在无偏估计的条件下，估计量的方差越小，估计就越有效。

定义 5.20　对同一个总体的两个无偏估计量 $\hat{\theta}_1$ 和 $\hat{\theta}_2$，若 $D(\hat{\theta}_1) < D(\hat{\theta}_2)$，则称 $\hat{\theta}_1$ 是比 $\hat{\theta}_2$ 更有效的一个估计量。

证明：总体均值 μ 的两个无偏估计量 $\bar{x}$ 和 $\bar{x}'$，$\bar{x}$ 比 $\bar{x}'$ 更有效。

$$
\begin{aligned}
D(\bar{x}) &= D(\frac{x_1 + x_2 + \cdots + x_n}{n}) \\
&= \frac{1}{n^2}\sum D(x_i) = \frac{1}{n^2} \times n\sigma^2 = \frac{1}{n}\sigma^2 \\
D(\bar{x}') &= D(\frac{a_1 x_1 + a_2 x_2 + \cdots + a_n x_n}{a_1 + a_2 + \cdots + a_n}) \\
&= \frac{\sum a_i^2 \sigma^2}{(\sum a_i)^2} = \frac{\sum a_i^2}{(\sum a_i)^2}\sigma^2
\end{aligned}
$$

现在比较 $\frac{1}{n}$ 和 $\frac{\sum a_i^2}{(\sum a_i)^2}$ 大小。

其中

$$
\begin{aligned}
(\sum a_i)^2 = \sum a_i^2 &+ 2a_1 a_2 + 2a_1 a_3 + \cdots + 2a_1 a_n \\
&+ 2a_2 a_3 + 2a_2 a_4 + \cdots + 2a_2 a_n \\
&+ 2a_3 a_4 + \cdots + 2a_3 a_n \\
&+ \cdots \\
&+ 2a_{n-1} a_n
\end{aligned}
$$

又有 $2a_i a_j \leqslant a_i^2 + a_j^2$，

所以 $(\sum a_i)^2 \leqslant n\sum a_i^2$，$\frac{\sum a_i^2}{(\sum a_i)^2} \geqslant \frac{1}{n}$，即 $\bar{x}$ 比 $\bar{x}'$ 更有效。

3. 一致性

定义 5.21　随着样本容量的增大，点估计量的值越来越接近被估计总体的参数，称为一致性。

例如，研究估计某个班级统计成绩的方差，该研究总体是这个班级，共有人数为 50 人。分别让 4 个不同的研究人员去估计，估计的结果如下。

第 1 个人，随机抽取了 20 人，得出的方差为 5.6。

第 2 个人，随机抽取了 26 人，得出的方差为 4.9。

第 3 个人，随机抽取了 30 人，得出的方差为 5.5。

第 4 个人，由于手中掌握更多的数据，随机抽取了 48 人，得出的方差为 5.2。

试问，哪个人得出的估计值最接近被估计的总体参数？答案是第 4 个人得出的估计值最接近总体的参数方差。因为他抽取的样本最大。

5.3.4 参数估计的思路

前面已经介绍过一个具体的点估计值无法给出点估计的可靠性，因此不能完全依赖于一个点估计值，而是应该围绕点估计值构造出总体参数的一个区间，即区间估计，而区间估计推断出总体的参数 $\mu,\pi,\sigma^2 p$ 的范围理论依据是样本的抽样分布。所以，估计总体均值 μ 要从样本均值 $\bar{x}$ 的抽样分布入手；估计总体方差 σ^2 要从样本方差 s^2 的抽样分布入手；估计总体比例 π 要从样本比例 p 的抽样分布入手。

区间估计就是在一定的置信水平下，得出总体参数的置信区间。曾经在概率中学习过当一个统计量服从标准正态分布、χ^2 分布、t 分布和 F 分布时，可以查表得出该统计量小于某个值所对应的概率，相反，给出某个概率，可以查表得出某个值，即 $P(X \leqslant x)$。给出上式的 x 的数值，由随机变量 X 的分布表中可得出概率 P。同样，先给出概率 P，可以由随机变量的分布表得出 x 数值。

根据以上方法，得到区间估计的思路：估计总体参数的区间范围，要从样本的抽样分布入手，同时把抽样分布用各种方法使其服从标准正态分布、χ^2 分布、t 分布和 F 分布 4 种分布之一。根据此思路，下面讲解一个总体参数的区间估计。

5.4 一个总体的参数区间估计

本节将介绍如何用样本统计量来构造总体均值的置信区间。

5.4.1 总体均值的区间估计

对总体均值进行区间估计时，需要考虑总体是否为正态总体、总体方差是否已知、用于构造估计量的样本是大样本还是小样本等几种情况。

1. 正态总体、方差已知，或非正态总体、大样本

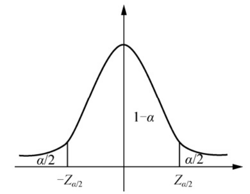
当总体服从正态分布且 σ 已知，或者总体不是正态分布但为大样本时，样本均值 $\bar{x}$ 的抽样分布均为正态分布，其数学期望为总体均值 μ，方差为 $\frac{1}{n}\sigma^2$，即 $\bar{x} \sim N(\mu,\frac{1}{n}\sigma^2)$。样本均值经过标准化以后的随机变量服从标准正态分布，即

$$Z=\frac{\bar{x}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$$

在置信水平 $1-\alpha$ 有

$$P(a \leqslant Z \leqslant b)=1-\alpha$$

根据标准正态分布的性质可得出



即 $a=-Z_{\alpha/2}$，$b=Z_{\alpha/2}$，所以有 $-Z_{\alpha/2}\leqslant Z\leqslant Z_{\alpha/2}$，又因为 $Z=\dfrac{\bar{x}-\mu}{\sigma/\sqrt{n}}$，有下式成立：

$$-Z_{\alpha/2}\leqslant Z=\frac{\bar{x}-\mu}{\sigma/\sqrt{n}}\leqslant Z_{\alpha/2}$$

由于总体方差是已知的，分母乘到两边，得

$$-Z_{\alpha/2}\,\sigma/\sqrt{n}\leqslant\bar{x}-\mu\leqslant Z_{\alpha/2}\,\sigma/\sqrt{n}$$

上式中，样本均值 $\bar{x}$ 可以利用样本数据计算出来，所以由上式可以得到总体均值 μ，即区间范围如下：

$$\bar{x}-Z_{\alpha/2}\,\sigma/\sqrt{n}\leqslant\mu\leqslant\bar{x}+Z_{\alpha/2}\,\sigma/\sqrt{n} \tag{5.10}$$

或者写成：

$$\bar{x}\pm Z_{\alpha/2}\,\sigma/\sqrt{n}$$

式中，$\bar{x}-Z_{\alpha/2}\,\sigma/\sqrt{n}$ 称为置信下限，$\bar{x}+Z_{\alpha/2}\,\sigma/\sqrt{n}$ 称为置信上限；$1-\alpha$ 称为置信水平；α 是事先所确定的一个概率值，也被称为风险值，它是总体均值不包括在置信区间的概率；$Z_{\alpha/2}$ 是标准正态分布上侧面积为 $Z_{\alpha/2}$ 时的 Z 值；$Z_{\alpha/2}\,\sigma/\sqrt{n}$ 是估计总体均值时的边际误差，也称估计误差或误差范围。这就是说，总体均值的置信区间由两部分组成：点估计值和描述估计量精度的误差值，该值称为边际误差。

2. 正态总体、方差未知，或非正态总体，但样本是大样本

1) 区间估计

如果总体服从正态分布且 σ 未知，或总体并不服从正态分布，但只要是在大样本条件下，样本均值 $\bar{x}$ 同样服从正态分布，即 $\bar{x}\sim N(\mu,\frac{1}{n}\sigma^2)$，经过标准化以后的随机变量还是服从标准正态分布，即

$$Z=\frac{\bar{x}-\mu}{\sigma/\sqrt{n}}\sim N(0,1)$$

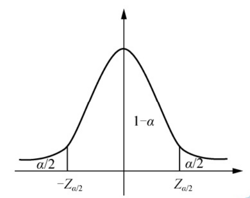
但此时的标准正态分布的统计量中包含了两个未知参数(μ 和 σ)，所以无法求出 μ 的区间，但由于样本是大样本，前面学习过一致性，随着样本容量的不断增加，样本所计算出的样本统计量非常接近于总体参数，所以这时可以用样本方差 s^2 代替总体方差 σ^2。所以有

$$Z=\frac{\bar{x}-\mu}{s/\sqrt{n}}\sim N(0,1)$$

在置信水平 $1-\alpha$ 有

$$P(a \leqslant Z \leqslant b) = 1-\alpha$$

根据标准正态分布的性质可得出



即 $a=-Z_{\alpha/2}$， $b=Z_{\alpha/2}$，所以有 $-Z_{\alpha/2} \leqslant Z \leqslant Z_{\alpha/2}$，又因为 $Z=\dfrac{\bar{x}-\mu}{s/\sqrt{n}}$，有下式成立：

$$-Z_{\alpha/2} \leqslant Z=\frac{\bar{x}-\mu}{s/\sqrt{n}} \leqslant Z_{\alpha/2}$$

由于总体方差是已知的，分母乘到两边，得到：

$$-Z_{\alpha/2}\, s/\sqrt{n} \leqslant \bar{x}-\mu \leqslant Z_{\alpha/2}\, s/\sqrt{n}$$

上式中，样本均值 $\bar{x}$ 可以利用样本数据计算出来，所以由上式可以得到总体均值 μ，即区间范围如下：

$$\bar{x}-Z_{\alpha/2}\, s/\sqrt{n} \leqslant \mu \leqslant \bar{x}+Z_{\alpha/2}\, s/\sqrt{n} \tag{5.11}$$

或者写成

$$\bar{x} \pm Z_{\alpha/2}\, s/\sqrt{n}$$

式中，$\bar{x}-Z_{\alpha/2}\, s/\sqrt{n}$ 称为置信下限，$\bar{x}+Z_{\alpha/2}\, s/\sqrt{n}$ 称为置信上限；这时 $Z_{\alpha/2}\, s/\sqrt{n}$ 是估计总体均值时的边际误差。

2) Excel 中的统计函数

利用 Excel 中的 NORMSINV 函数可以计算给定置信水平下的正态分布的分位数值。在 95%的置信水平下，相应的 $\alpha/2=0.025$。求 $Z_{\alpha/2}$ 的具体步骤如下。

第一步：进入 Excel 表格界面，单击“插入函数”按钮，弹出“插入函数”对话框，在对话框中单击“或选择类别”的下拉按钮，在弹出的下拉列表中选择“统计”选项，并在“选择函数”列表中选择 NORM.S.INV 选项，单击“确定”按钮，弹出“函数参数”对话框。

第二步：在“函数参数”对话框中的 Probability 文本框中输入“0.025”，得到“−1.95996”。

3. 正态总体、方差未知，样本是小样本

1) 区间估计

如果总体方差 σ^2 已知，而且是在小样本的情况下，也可以用样本方差 s^2 代替 σ^2，但此时样本样本均值经过标准化以后的随机变量服从自由度为$(n-1)$的 t 分布，其过程如下。

当总体服从正态分布，无论大小样本，样本均值 $\bar{x}$ 的抽样分布均为正态分布，即

$\bar{x} \sim N(\mu, \frac{1}{n}\sigma^2)$。样本均值经过标准化以后的随机变量服从标准正态分布：

$$Z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$$

此时该统计量中包含了两个未知参数(μ和σ)，所以无法求出μ的区间。又是小样本，不能直接用样本方差s^2代替σ^2，此时从标准正态无法得出总体μ的区间估计。若用χ^2分布，χ^2是标准正态分布的平方加和，经过平方加和后，其式中还有两个未知参数(μ和σ)，还是不可以。此时可用t分布。同时有

$$\chi^2 = \frac{(n-1)s^2}{\sigma^2} \sim \chi^2(n-1)$$

所以有

$$t = \frac{Z}{\sqrt{\chi^2/(n-1)}} \sim t(n-1)$$

把上两式代入得到：

$$t = \frac{\dfrac{\bar{x} - \mu}{\sigma/\sqrt{n}}}{\sqrt{\dfrac{(n-1)s^2}{\sigma^2}/(n-1)}} \sim t(n-1)$$

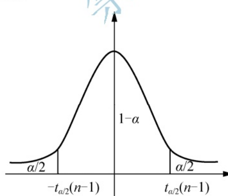
整理后，得到下式：

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}} \sim t(n-1)$$

在置信水平$1-\alpha$有

$$P(a \leqslant t \leqslant b) = 1-\alpha$$

根据标准正态分布的性质可得出



即$a = -t_{\alpha/2}(n-1)$，$b = t_{\alpha/2}(n-1)$，所以有$-t_{\alpha/2}(n-1) \leqslant t \leqslant t_{\alpha/2}(n-1)$，又因为$t=\frac{\bar{x} - \mu}{s/\sqrt{n}}$，有下式成立：

$$-t_{\alpha/2}(n-1) \leqslant t = \frac{\bar{x} - \mu}{s/\sqrt{n}} \leqslant t_{\alpha/2}(n-1)$$

经整理，得到总体均值μ的区间，即区间范围如下：

$$\bar{x} - t_{\alpha/2}(n-1)s/\sqrt{n} \leqslant \mu \leqslant \bar{x} + t_{\alpha/2}(n-1)s/\sqrt{n} \tag{5.12}$$

或者写成：

$$\bar{x} \pm t_{\alpha/2}(n-1)s/\sqrt{n}$$

式中，$t_{\alpha/2}(n-1)s/\sqrt{n}$ 为估计总体均值时的边际误差。

2) Excel 中的统计函数

利用 Excel 中的 TINV 函数可以计算给定置信水平下的 t 分布的临界值。设自由度 $\text{df}=15$，在 95%的置信水平下，相应的 $\alpha/2=0.025$。求得具体步骤如下。

第一步：进入 Excel 表格界面，单击“插入函数”按钮，弹出“插入函数”对话框，在对话框中单击“或选择类别”的下拉按钮，在弹出的下拉列表中选择“统计”选项，并在“选择函数”列表中选择 T.INV 选项，单击“确定”按钮，弹出“函数参数”对话框。

第二步：在“函数参数”对话框中的 Probability 文本框中输入“0.05”，在 Deg_freedom 文本框中输入“15”，该函数自动返回 $Z_{\alpha/2}$ 的值为“2.131449536”。

将以上总体均值的区间估计进行总结，见表 5-4 所示。

表 5-4 不同情况下总体均值的区间估计

总体分布	样本容量	σ 已知	σ 未知
正态分布	大样本($n>30$)	$\bar{x}\pm Z_{\alpha/2}\sigma/\sqrt{n}$	$\bar{x}\pm Z_{\alpha/2}s/\sqrt{n}$
	小样本($n\leqslant 30$)	$\bar{x}\pm Z_{\alpha/2}\sigma/\sqrt{n}$	$\bar{x}\pm t_{\alpha/2}s/\sqrt{n}$
非正态分布	大样本($n>30$)	$\bar{x}\pm Z_{\alpha/2}\sigma/\sqrt{n}$	$\bar{x}\pm Z_{\alpha/2}s/\sqrt{n}$

【例 5.3】一家罐装饮料生产企业，要求每罐的平均容量为 255 mL，标准差为 5mL，为了对产品质量进行监测，从某天生产的一批产品中随机抽取 40 罐进行研究，测得每罐的平均容量为 255.9 mL。已知产品容量的分布服从正态分布。试估计该批产品平均重量的置信区间，置信水平为 95%。

解：已知总体的标准差 $\sigma=5$，所以无论样本为大小样本，对总体均值进行区间估计，使用式(5.10)：

$$\bar{x}-Z_{\alpha/2}\,\sigma/\sqrt{n}\leqslant\mu\leqslant\bar{x}+Z_{\alpha/2}\,\sigma/\sqrt{n}$$

其中 $n=40$，置信水平为 $1-\alpha=95\%$，查标准正态分布表得 $Z_{\alpha/2}=1.96$，同时有 $\bar{x}=255.9$。所以有

$$255.9\pm1.96\times\frac{5}{\sqrt{40}}$$

$$254.4\leqslant\mu\leqslant257.4$$

【例 5.4】 一家保险公司想知道某险种被保人的年龄范围，从而评估该险种的盈利情况，收集到由 36 位投保人组成的随机样本，得到每位被保险人的年龄数据，见表 5-5 所示。

表 5-5 36 位被保险人的年龄

23	35	40	25	35	42	36	41	46
40	32	35	40	45	48	40	32	48
34	23	27	40	41	38	30	42	28
44	43	37	21	38	28	35	41	38

试建立被保险人年龄99%的置信区间。

解：根据题意，总体方差未知，但$n=36$为大样本，要估计总体均值的区间，所以使用式(5.11)：

$$\bar{x}-Z_{\alpha/2}s/\sqrt{n}\leqslant\mu\leqslant\bar{x}+Z_{\alpha/2}s/\sqrt{n}$$

根据样本的数据，计算样本的均值和样本方差如下：

$$\bar{x}=\frac{\sum x_i}{n}=36.4 \qquad s=\sqrt{\frac{\sum(x_i-\bar{x})^2}{n-1}}=7.14$$

置信水平为$1-\alpha=99\%$，查标准正态分布表得$Z_{\frac{\alpha}{2}}=2.58$。

所以有

$$36.4\pm2.58\times\frac{7.14}{\sqrt{36}}$$

$$33.33\leqslant\mu\leqslant39.47$$

【例 5.5】 一家食品生产企业以生产袋装食品为主，每天产量大约为 8 000 袋。按规定，每袋食品的重量应为 100g。为对产品质量进行监测，企业质检部经常要进行抽检，以分析每袋重量是否符合要求。现从某天生产的一批食品中随机抽取了 25 袋，测得每袋重量见表 5-6 所示。

表 5-6 25 袋食品的重量

单位：g

112.5	101.0	103.0	102.0	100.5	102.6	107.5	95.0	108.8
115.6	100.0	95.4	102.0	101.6	102.2	116.6	123.5	97.8
108.6	105.0	136.8	102.8	101.5	95.4	93.3	—	—

已知产品重量的分布服从正态分布。试估计该批产品平均重量的置信区间，置信水平为95%。

解：总体方差未知，置信水平$1-\alpha=95\%$，查标准正态分布表得$t_{\alpha/2}(24)=2.39$。根据样本数据计算的样本均值和样本标准差为

$$\bar{x}=\frac{\sum_{i=1}^{n}x_i}{n}=\frac{2634}{25}=105.36 \qquad s=\sqrt{\frac{\sum(x_i-\bar{x})^2}{n-1}}=\sqrt{95.31}=9.76$$

根据式(5.12)得：

$$\bar{x}-t_{\alpha/2}(n-1)s/\sqrt{n}\leqslant\mu\leqslant\bar{x}+t_{\alpha/2}(n-1)s/\sqrt{n}$$

$$105.36\pm2.39\times\frac{9.76}{\sqrt{25}}$$

最后的区间范围为

$$100.69\leqslant\mu\leqslant110.03$$

5.4.2 总体比例的区间估计

本节只讨论大样本情况下总体比例的估计问题，当样本容量足够大时，比例p的抽样

分布可用正态分布近似，即

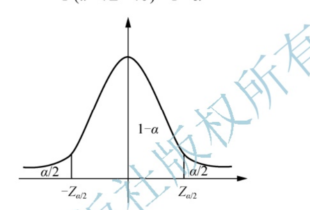
$$p \sim N(\pi, \frac{\pi(1-\pi)}{n})$$

样本比例经标准化以后的随机变量服从标准正态分布，即

$$z = \frac{p-\pi}{\sqrt{\frac{\pi(1-\pi)}{n}}} \sim N(0,1)$$

与总体均值的区间估计类似，在置信水平$1-\alpha$有

$$P(a \leqslant Z \leqslant b) = 1-\alpha$$



即$a = -Z_{\alpha/2}$，$b = Z_{\alpha/2}$，所以有$-Z_{\alpha/2} \leqslant Z \leqslant Z_{\alpha/2}$，又因为$z = \frac{p-\pi}{\sqrt{\frac{\pi(1-\pi)}{n}}}$，有下式成立：

$$-Z_{\alpha/2} \leqslant Z = \frac{p-\pi}{\sqrt{\frac{\pi(1-\pi)}{n}}} \leqslant Z_{\alpha/2}$$

$$p - Z_{\alpha/2}\sqrt{\frac{\pi(1-\pi)}{n}} \leqslant \pi \leqslant p + Z_{\alpha/2}\sqrt{\frac{\pi(1-\pi)}{n}}$$

又因为是大样本，所以用样本的统计量p代替两边的π，得到区间范围如下：

$$p - Z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}} \leqslant \pi \leqslant p + Z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}} \qquad (5.13)$$

或者写成：

$$p \pm Z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}$$

其中$Z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}$是估计总体比例时的边际误差。

【例 5.6】 某城市想要估计下岗职工中女性所占的比例，所以随机抽取了 1 000 个下岗职工，其中 650 人为女性职工。试以 90%的置信水平估计该城市下岗职工中女性比例的置信区间。

解：已知 n=1 000，置信水平$1-\alpha = 90\%$，查标准正态分布表得$Z_{\alpha/2} = 1.645$。根据样本数据计算的样本比例为

$$p = \frac{650}{1\,000} = 65\%$$

根据式(5.13)得：

$$p \pm Z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}} = 65\% \pm 1.96 \times \sqrt{\frac{65\% \times (1-65\%)}{1\,000}}$$

即 $65\% \pm 1.5\% = (63.5\%, 66.5\%)$。

5.4.3 总体方差的区间估计

1. 区间估计

本节只讨论正态总体方差的区间估计问题。构造总体方差的区间估计，要从样本方差 s^2 入手，由于样本方差 $\frac{(n-1)s^2}{\sigma^2} \sim \chi^2(n-1)$，因此用 χ^2 分布构造总体方差的置信区间。

$$\chi^2 = \frac{(n-1)s^2}{\sigma^2} \sim \chi^2(n-1)$$

在置信水平 $1-\alpha$ 有

$$P(a \leqslant \chi^2 \leqslant b) = 1-\alpha$$

所以有

$$\chi^2_{1-\alpha/2}(n-1) \leqslant \chi^2 \leqslant \chi^2_{\alpha/2}(n-1)$$

由于 $\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$，因此可以用它来代替 χ^2，于是有

$$\chi^2_{1-\alpha/2}(n-1) \leqslant \frac{(n-1)s^2}{\sigma^2} \leqslant \chi^2_{\alpha/2}(n-1)$$

最后可以推导出总体方差 σ^2 在 $1-\alpha$ 置信水平下的置信区间为

$$\frac{(n-1)s^2}{\chi^2_{\alpha/2}(n-1)} \leqslant \sigma^2 \leqslant \frac{(n-1)s^2}{\chi^2_{1-\alpha/2}(n-1)} \tag{5.14}$$

2. Excel 中的统计函数

利用 Excel 中的 CHIINV 函数可以计算给定置信水平下的 χ^2 分布的分位数值。设自由度 df = 24，在 95%的置信水平下，相应的 $\alpha/2$=0.025。求 $\chi^2_{\alpha/2}$ 的具体步骤如下。

第一步：进入 Excel 表格界面，单击“插入函数”按钮，弹出“插入函数”对话框，在对话框中单击“或选择类别”的下拉按钮，在弹出的下拉列表中选择“统计”选项，并在“选择函数”列表中选择 CHIINV 选项，单击“确定”按钮，弹出“函数参数”对话框。

第二步：在“函数参数”对话框中的 Probability 文本框中输入“0.025”，在 Deg_freedom

文本框中输入“24”，该函数自动返回 $\chi^2_{\alpha/2}$ 的值为“39.36407706”。同样可得到 $\chi^2_{1-\alpha/2}$ =0.975 的值为 “12.40115026”。

【例 5.7】 仍利用例 5.5 的数据，以 95%的置信水平建立该种食品重要的方差的置信区间。

解：已知根据样本的数据计算出 $s=\sqrt{\dfrac{\sum(x_i-\overline{x})^2}{n-1}}=\sqrt{95.31}=9.76$，置信水平 $1-\alpha=95\%$，则查表可得：

$$\chi^2_{0.975}(24)=12.401\,15$$

$$\chi^2_{0.025}(24)=39.364\,08$$

$$\frac{(n-1)s^2}{\chi^2_{\alpha/2}(n-1)}\leqslant\sigma^2\leqslant\frac{(n-1)s^2}{\chi^2_{1-\alpha/2}(n-1)}$$

$$\frac{24\times9.76^2}{39.364\,08}\leqslant\sigma^2\leqslant\frac{24\times9.76^2}{12.401\,15}$$

$$58.08\leqslant\sigma^2\leqslant184.35$$

5.5 样本容量的确定

在进行参数估计之前，首先应确定一个适当的样本容量，也就是应该抽取一个多大的样本来估计总体参数。在进行参数区间估计时，一是希望提高估计的可靠程度，二是提高区间估计的精确性。但在样本容量一定时，两者往往是对立的。

例如，要说出某一天会下雨，置信区间并不宽，但是可靠性相对较低，如果说第三季度会下一场雨，尽管很可靠，但准确性又太差，也就是置信区间太宽的估计是没有意义的。如果既想缩小置信区间，又不想降低置信程度，就需要增加样本容量，但样本容量的增加也会受到许多限制，如会增加调查的费用和工作量。通常来说，样本容量的确定与可容忍的置信区间的宽度及对此区间设置的置信水平有一定关系。因此，如何确定一个适当的样本容量，也是抽样估计中需要考虑的问题。

5.5.1 估计总体均值时样本容量的确定

如前所述，总体均值的置信区间是由样本均值 $\overline{x}$ 和边际误差两部分组成。

1. 总体方差已知

在重复抽样或无限总体抽样的条件下，总体方差是已知时，边际误差为 $Z_{\alpha/2}\,\sigma/\sqrt{n}$。$Z_{\alpha/2}$ 的值和样本容量 n 共同确定了边际误差的大小。也就是说，一旦确定了置信水平 $1-\alpha$，那么 $Z_{\alpha/2}$ 的值就确定了。根据给定的 $Z_{\alpha/2}$ 的值和总体标准差 σ，就可以确定任一希望的边际误差内所需要的样本容量。令 E 代表希望达到的边际误差，即

$$Z_{\alpha/2}\,\sigma/\sqrt{n}\leqslant E$$

由此可以推导出确定样本容量的公式为

$$n \geqslant \frac{(Z_{\alpha/2})^2 \sigma^2}{E^2} \tag{5.15}$$

2. 总体方差未知，大样本

相同的原理，当总体方差未知时，大样本时，可得出样本容量的公式为

$$n \geqslant \frac{(Z_{\alpha/2})^2 s^2}{E^2} \tag{5.16}$$

3. 总体方差未知，小样本

总体方差未知，小样本时，样本容量的公式为

$$n \geqslant \frac{[t_{\alpha/2}(n-1)]^2 s^2}{E^2} \tag{5.17}$$

从上面几个公式可以看出，样本容量与置信水平成正比，在其他条件不变的情况下，置信水平越大，所需的样本容量也越大；样本容量与总体方差成正比，总体的差异越大，所要求的样本容量也越大；样本容量与边际误差的平方成反比，即可以接受的边际误差的平方越大，所需的样本容量就越小。

注意：计算出的样本容量不一定是整数，通常是将样本容量取成最小的整数，也就是将小数点后面的数值一律进位成整数，如 25.68 取 26，25.01 也取 26，即这就是样本容量的圆整法则。

【例 5.8】 某超市想要估计每个顾客平均每次购物花费的金额。根据过去的经验，标准差为 120 元，现要求以 90%的置信水平估计每个顾客平均购物金额的置信区间，并要求边际误差不超过 20 元，应最少抽取多少个顾客作为样本？

解：根据题意可知，想要估计总体均值，且总体方差是已知，所以使用式(5.15)，有

$$n \geqslant \frac{(Z_{\alpha/2})^2 \sigma^2}{E^2} = \frac{(1.645 \times 120)^2}{20^2} = 97.4169$$

即最少要抽取 98 个顾客作为样本。

5.5.2 估计总体比例时样本容量的确定

与估计总体均值时样本容量的确定方法类似，在重复抽样或无限总体抽样的条件下，估计总体比例置信区间的边际误差为 $Z_{\alpha/2}\sqrt{\frac{\pi(1-\pi)}{n}}$。

$$Z_{\alpha/2}\sqrt{\frac{\pi(1-\pi)}{n}} \leqslant E$$

由此可以推导出重复抽样或无限总体抽样条件下确定样本容量的公式为

$$n \geqslant \frac{(Z_{\alpha/2})^2 \pi(1-\pi)}{E^2} \tag{5.18}$$

在实际应用中，如果不知道 π 的值，可以通过类似的样本比例来代替。

【例 5.9】 根据以往的生产统计，某产品的合格率约为 95%，现要求边际误差为 4%，

在求 90%的置信区间时，应最少抽取多少个产品作为样本？

解：已知 $\pi = 95\%$，$E = 4\%$，$Z_{\alpha/2} = 1.645$，根据式(5.17)得：

$$n \geqslant \frac{(Z_{\alpha/2})^2 \pi(1-\pi)}{E^2} = \frac{(1.645)^2 \times 0.95 \times (1-0.95)}{0.04^2} = 80.335$$

即应最少抽取 81 个产品作为样本。

习　　题

一、填空题

1. 样本抽样分布是指样本的(　　)分布。
2. 当总体是正态分布时，无论样本量的大小，样本均值 $\bar{x}$ 都服从(　　)。
3. 总体(或样本)中具有某种属性的单位数与全部单位数的比值称为(　　)。
4. 估计总体参数 θ 的统计量的名称，称为(　　)。
5. 估计总体参数时计算出来的估计量的具体数值，称为(　　)。
6. 用样本统计量 $\hat{\theta}$ 的某个取值直接作为总体参数 θ 的估计值，称为参数的(　　)。
7. 在点估计值的基础上，给出总体参数估计的一个范围，称为参数的(　　)。
8. 由样本统计量所构造的总体参数的估计区间，称为(　　)，其中区间的最小值称为(　　)，最大值称为(　　)。
9. 如果将构造置信区间的步骤重复多次，那么区间中包含总体参数真值的次数所占的比率，称为(　　)。
10. 评价估计量的标准有(　　)、(　　)、(　　)。
11. 随着样本容量的增大，点估计量的值越来越接近被估计总体的参数，这种评价标准称为(　　)。
12. 当总体服从正态分布且 σ 已知，或者总体不是正态分布但为大样本时，样本均值 $\bar{x}$ 的抽样分布均为(　　)。
13. 总体均值的置信区间等于样本均值加减(　　)。
14. 总体参数的置信区间是由样本统计量的(　　)加减(　　)得到的。
15. 其他条件不变的情况下，90%的置信区间比 95%的置信区间(　　)。

二、单项选择题

1. 当总体是非正态分布，样本容量为大样本时，样本均值 $\bar{x}$ 服从(　　)。
   A. 正态分布　　B. 标准正态分布　　C. t 分布　　D. χ^2 分布
2. 当样本均值 $\bar{x}$ 的抽样分布服从正态分布时，其分布的均值为(　　)。
   A. $\bar{x}$　　B. μ/n　　C. μ　　D. σ^2/n
3. 当样本均值 $\bar{x}$ 的抽样分布服从正态分布时，其分布的方差为(　　)。
   A. $\sigma/\sqrt{n}$　　B. $\sigma^2/\sqrt{n}$　　C. μ/n　　D. σ^2/n
4. 在大样本的情况下，可推倒出样本比例 p 的抽样分布为(　　)。
   A. 正态分布　　B. 标准正态分布　　C. t 分布　　D. χ^2 分布
5. 在大样本的情况下，在重复抽样条件下，样本比例分布的方差为(　　)。
   A. π　　B. $1-\pi$　　C. $\pi(1-\pi)$　　D. $\pi(1-\pi)/n$
6. 样本方差所有可能取值形成的概率分布为(　　)。
   A. 正态分布　　B. 标准正态分布　　C. t 分布　　D. χ^2 分布

7. 统计量的标准误差是指(　　)。

A. 样本观测值的标准差　　B. 总体观测值的标准差

C. 样本统计量抽样分布的标准差　　D. 总体统计量的标准差

8. 下列说法中不正确的是(　　)。

A. 样本均值是总体均值的点估计

B. 样本比例是总体比例的点估计

C. 如果抽样分布的均值等于总体参数，则该统计量称作参数的无偏估计

D. 如果抽样分布的均值不等于总体参数，则该统计量称作参数的无偏估计

9. 对同一个总体的两个无偏估计量 $\hat{\theta}_1$ 和 $\hat{\theta}_2$，若称 $\hat{\theta}_1$ 是比 $\hat{\theta}_2$ 更有效的一个估计量，则 $\hat{\theta}_1$ 和 $\hat{\theta}_2$ 需要满足的条件是(　　)。

A. $D(\hat{\theta}_1) > D(\hat{\theta}_2)$　　B. $D(\hat{\theta}_1) < D(\hat{\theta}_2)$

C. $\mathrm{COV}(\hat{\theta}_1) > \mathrm{COV}(\hat{\theta}_2)$　　D. $\mathrm{COV}(\hat{\theta}_1) < \mathrm{COV}(\hat{\theta}_2)$

10. 从服从正态总体的总体中抽取容量为 8、24、64 的样本，随着样本容量的增大，样本均值的标准差(　　)。

A. 增加　　B. 减少　　C. 保持不变　　D. 服从 χ^2 分布

11. 总体均值为 20，标准差为 6，从此总体中随机抽取容量为 36 的样本，则样本均值和抽样分布的标准误差分别为(　　)。

A. 20, 6　　B. 20, 1　　C. 20, 4　　D. 6, 6

12. 某大学附近的一家超市记录了过去两年每天的销售额，其每天销售额的均值为 2 000 元，标准差为 500 元。由于在某些节日的销售额偏高，所以每日销售额的分布时右偏的。假设从这两年中随机抽取 100 天，并计算这 100 天的平均销售额，则样本均值的抽样分布是(　　)。

A. 正态分布，均值为 200 元，标准差为 50 元

B. 正态分布，均值为 2 000 元，标准差为 50 元

C. 右偏，均值为 2 000 元，标准差为 500 元

D. 正态分布，均值为 2 000，标准差为 500 元

13. 从均值为 2 500，标准差为 500 的总体中抽取容量为 100 的简单随机样本，用样本均值估计总体均值。样本均值的数学期望是(　　)。

A. 3 000　　B. 2 500　　C. 2 000　　D. 1 500

14. 从均值为 2 500，标准差为 500 的总体中抽取容量为 100 的简单随机样本，用样本均值估计总体均值。样本均值的标准差是(　　)。

A. 100　　B. 150　　C. 50　　D. 500

15. 假定总体比例为 0. 8，从此总体中抽取容量为 100 的样本，则样本比例的数学期望为(　　)。

A. 0.8　　B. 0.4　　C. 0.04　　D. 0.001 6

16. 假定总体比例为 0. 8，从此总体中抽取容量为 100 的样本，则样本比例的标准差为(　　)。

A. 0.8　　B. 0.4　　C. 0.04　　D. 0.001 6

17. 估计一个正态总体方差时，应使用的分布是(　　)分布。

A. 标准正态　　B. F　　C. t 分布　　D. χ^2

18. 当总体的方差未知，且为大样本的情况，对总体均值进行估计，所使用的分布是(　　)分布。

A. 标准正态　　B. F　　C. t　　D. χ^2

19. 当总体的方差未知，且为小样本的情况，对总体均值进行估计，所使用的分布是(　　)分布。

A. 标准正态　　B. F　　C. t　　D. χ^2

20. 当总体的方差已知，且为大样本的情况，对总体均值进行估计，所使用的分布是(　　)分布。

A. 标准正态　　B. F　　C. t　　D. χ^2

21. 当总体的方差已知，且为大样本的情况，对总体均值进行估计，所使用的分布是(　　)分布。

A. 标准正态　　B. F　　C. t　　D. χ^2

22. 在进行参数估计时，评价估计量的标准之一是使估计量抽样分布的数学期望等于被估计的总体参数，这一评价标准称为(　　)。

A. 充分性　　B. 无偏性　　C. 有效性　　D. 一致性

23. 在进行参数估计时，评价估计量的标准之一是使它与总体参数的离差越小越好，这一评价标准称为(　　)。

A. 充分性　　B. 无偏性　　C. 有效性　　D. 一致性

24. 根据某班统计学成绩的一个样本，估计全班同学统计学平均成绩的 95%的置信区间为 75～85 分。则全班同学统计学的平均分数(　　)。

A. 有 95%的概率落在这个区间内　　B. 有 5%的概率落在这个区间内

C. 一定落在这个区间内　　D. 可能在这一区间内，也可能不在这一区间内

25. 当置信水平一定时，置信区间的宽度(　　)。

A. 同样本容量的大小无关　　B. 同样本容量的平方根成正比

C. 随样本容量的增大而减小　　D. 随样本容量的增大而增大

26. 当样本容量一定时，置信区间的宽度(　　)。

A. 同置信水平的大小无关　　B. 同置信水平的平方成正比

C. 随置信水平的增大而减小　　D. 随置信水平的增大而增大

27. 在置信水平一定的情况下，容量大的样本比容量小的样本所构造的置信区间(　　)。

A. 可能宽可能窄　　B. 相同

C. 要窄　　D. 要宽

28. 在重复抽样或无限总体抽样的条件下，总体方差是已知时，边际误差为(　　)。

A. $Z_{\alpha/2}\sigma/\sqrt{n}$　　B. $Z_{\alpha/2}\sigma^2/n$　　C. $Z_{\alpha}\sigma/\sqrt{n}$　　D. $Z_{\alpha}\sigma^2/n$

29. 在重复抽样或无限总体抽样的条件下，总体方差是未知，大样本情况下，边际误差为(　　)。

A. $Z_{\alpha/2}s/\sqrt{n}$　　B. $t_{\alpha/2}(n-1)s/\sqrt{n}$　　C. $Z_{\alpha}s/\sqrt{n}$　　D. $t_{\alpha}(n-1)s/\sqrt{n}$

30. 在重复抽样或无限总体抽样的条件下，总体方差是未知，小样本情况下，边际误差为(　　)。

A. $Z_{\alpha/2}s/\sqrt{n}$　　B. $t_{\alpha/2}(n-1)s/\sqrt{n}$　　C. $Z_{\alpha}s/\sqrt{n}$　　D. $t_{\alpha}(n-1)s/\sqrt{n}$

三、计算题

1. 从均值为 81，标准差为 12 的总体中，抽取一个容量为 100 的简单随机样本，估计总体均值 μ 的置信区间。

(1) $\bar{x}$ 的数学期望是多少？

(2) $\bar{x}$ 的标准差为多少？

(3) 在 95%的置信水平下，边际误差是多少？

(4) 求总体均值的 95%的置信区间。

2. 从一个总体标准差为 4 的总体中抽取一个样本容量为 36 的样本，样本的均值为 20，则样本均值的抽样标准差为多少？

3. 从总体均值为 10，标准差为 100 的总体中，抽取一个样本容量为 20 $\bar{x}_{20}$ 的随机样本，样本均值为 $\bar{x}_{20}$，同样，再抽样一个样本容量为 50 的随机样本，样本均值为 $\bar{x}_{50}$，分布描述 $\bar{x}_{20}$ 和 $\bar{x}_{50}$ 的抽样分布。

4. 从总体比例 π=0.5 的总体中，抽取一个样本容量为 100 的随机样本。

(1) p 的数学期望是多少？

(2) p 的标准差为多少？

(3) 在 90%的置信水平下，边际误差是多少？

(4) 求总体比例 90%的置信区间。

5. 设总体比例为 0.8，分别从总体中抽取样本容量为 100、500、1 000 的样本。

(1) 分别计算每个样本比例的标准差。

(2) 随着样本容量的增大，样本比例的标准差是如何变化的？

6. 某研究机构想了解现在每个家庭每天看电视的平均时间，随机抽取了 200 个家庭所为研究对象，测得每个家庭每天看电视的平均时间为 6.25h，标准差为 2.5h，求现在每个家庭每天看电视平均时间的置信区间。置信水平分别为 90%、95%和 99%。

7. 某小学的班主任想了解班级学生上学从家到学校的距离，随机抽取了 16 名学生组成的一个样本，得到他们从家到学校的距离(km)如下：

10 20 15 9 12 15 21 11 16 18 12 16 13 8 10 15

求班上学生从家到学校平均距离 90 的置信区间。

8. 某品牌的灌装饮料，每瓶标准容量为 500mL，现从某天生产的一批产品中随机抽取 50 瓶进行检查，测得每瓶的容量见表 5-7 所示。

表 5-7 样本数据

每瓶容量/mL	瓶数
496～498	2
498～500	3
500～502	34
502～504	7
504～506	4
合计	50

试确定该品牌饮料平均容量 95%的置信区间。

9. 某高校提出了一项工资改革措施，为估计该校教师赞成这项改革措施的人数，从全校教职工中随机抽取了 100 人进行调查，其中赞成该项措施的人数占 20%，求总体比例的 90%和 95%的置信区间。

10. 在药品制造业，药品的重量非常关键，对某种特定的药物进行检查，从 25 个样本中得到样本的标准差为 0.6g，求该药物重量的总体方差 95%的置信区间。

11. 拥有工商管理学士学位的大学毕业生年薪的标准差大约为 2 000 元，假定想要估计年薪 95%的置信区间，并希望边际误差为 400 元，应抽取多大的样本容量？

12. 根据以往的生产经验，某种产品的合格率为 98%，如果要求该产品合格率 95%的置信区间，且要求边际误差不超过 5%，应抽取多大的样本容量？

第 6 章 假设检验

教学目标

1. 掌握假设检验的步骤。
2. 了解假设检验的基本问题。
3. 掌握一个总体参数(均值、比例和方差)的假设检验。
4. 掌握两个总体均值之差的假设检验。
5. 掌握两个总体比例之差的假设检验。
6. 掌握两个总体方差之比的假设检验。
7. 掌握假设检验的软件操作过程。

引入案例

女子体温一般比男子约高 0.3℃

当问起健康的成年人中女子体温是否与男子体温相同时，多数人的回答是不相同，但不知道是多少，根据有关数据显示，女子的体温比男子的高于 0.3℃，那么这个结论是否是正确的呢？表 6-1 是一个研究人员测量的 50 个健康成年人的体温（℃）数据。

表 6-1　50 个健康成年人体温测量数据表

男	36.2	36.9	36.2	36.1	37.1	37	36.6	36.1	36.7	36.8
	36.9	36.6	36.7	36.5	36.2	36.7	37.2	37.1	36.3	37
	36.9	36.2	37.3	36.6	36.3	36.9	36.4	37	—	—
女	37.3	37	37.1	36.7	36.9	36.5	37.5	37	36.6	37.2
	37.1	36.1	36.4	36.9	36.4	36.7	36.9	37.1	37.1	37.4
	36.7	37	—	—	—	—	—	—	—	—

根据样本数据计算的平均值为：男生的平均值为 36.7℃，标准差为 0.361 4℃；女子的体温平均值为 36.9℃，标准差为 0.349 0℃。从样本数据可得出女子的体温平均比男子的体温高于 0.2℃，那么我们是不是就可以得出女子的体温比男子的要高，而不是高于 0.3℃，是 0.2℃呢？本章的内容将提供一套标准统计程序来检验这个问题。

6.1 假设检验的基本理论

6.1.1 假设检验的定义

在现实生活中，人们经常要对某个“假设”做出判断，确定它的真假。在研究领域中，研究者在检验一种新的理论时，往往也是首先提出一种自己认为是正确的看法，即假设。而在统计学中，“假设”就是对总体参数的一种事先猜想。

定义 6.1　对总体参数的具体数值所做的陈述，称为假设，也称统计假设。

一个假设的提出总是以一定的理由为基础的，但这些理由通常又是不完全的，因而产生了“检验”的需求，也就是要进行判断。例如，在对某一品牌洗衣粉的抽检中，抽检人员需要判断其净含量是否达到了说明书中所声明的质量；公司在收到一批货物时，质检人员需要判断该批货物的属性是否与合同中规定的一致；某企业使用自动线生产产品，质检人员检验自动线生产是否正常等。

当提出假设后，通常要对假设进行判断，即假设检验。假设检验是利用样本信息判断假设是否成立的过程。

定义 6.2　先对总体参数提出某种假设，然后利用样本信息判断假设是否成立的过程，称为假设检验。

6.1.2 假设检验的基本步骤

1. 传统假设检验的基本步骤

传统假设检验的基本步骤有 4 个。

(1) 提出原假设 H_0 和备择假设 H_1。

(2) 构造检验的统计量，并计算其值。

(3) 根据给出的显著性水平 α，确定拒绝原假设 H_0 的区域。

(4) 统计决策。

下面一一来介绍每一步的内容。

1) 提出原假设 H_0 和备择假设 H_1

首先要清楚什么是原假设 H_0 和备择假设 H_1。

定义 6.3　通常将研究者想收集证据予以支持的假设称为备择假设，即研究人员认为正确的观点，用 H_1 表示。

备择假设通常是用于支持自己的看法。例如，质检部门要检验某车间某天生产的产品是否合格，就应该把他们认为的想法作为备择假设；我们正在做一项研究，并想使用假设检验来支持我们的说法，就应该把我们认为正确的看法作为备择假设。

定义 6.4　通常将研究人员想收集证据予以反对的假设称为原假设，即研究人员认为不正确的观点，用 H_0 表示。

在假设检验中，确定原假设和备择假设十分重要，它直接关系到检验的结论。从原假设和备择假设的定义来看，如果找出研究人员，之后再确定研究人员的想法，其予以支持

的观点就是备择假设，相反，就是他反对的观点，即为原假设。下面举例说明原假设和备择假设的建立过程。

【例 6.1】 一种食品生产企业以生产袋装食品为主，按规定每袋的标准净含量要求为 100g，为对生产过程进行控制，质量监测人员定期对袋装食品进行检查，以确定生产出来的食品是否符合要求。如果平均净含量大于 100g 或小于 100g，都表明生产过程不正常，必须进行调整。试陈述用来检验生产过程是否正常的原假设和备择假设。

解：如果企业生产的食品净含量 μ=100，表明生产过程正常；如果 $\mu>100$ 或 $\mu<100$，表明生产过程不正常。究竟哪个作为原假设，哪个作为备择假设，需要先找出研究者。根据题意可知，研究者为质量监测人员，如果他认为产品是合格的话，他不用去检查，只有他认为不合格，才会去检查。所以他认为的观点是 $\mu>100$ 或 $\mu<100$，即生产过程不正常，此观点即为备择假设，相反的为原假设。

所以研究者建立的原假设和备择假设应为

H_0： μ=100 (生产过程正常)

H_1： $\mu\neq100$ (生产过程不正常)

【例 6.2】 某品牌奶粉在产品说明书中声称：平均净含量不少于 400g。从消费者的利益出发，有关研究人员要通过抽检其中的一批奶粉产品来检验该产品制造商的说明是否属实。试陈述用于检验的原假设与备择假设。

解：该品牌奶粉平均净含量的真值为 μ。如果抽检的结果发现 $\mu<400$，则表明该产品说明书中关于其净含量的内容是不真实的，有关部门应对其采取相应的措施。

该题的研究人员是从消费者的利益出发，对产品说明书产生的质疑，否则不会去抽检产品，所以研究者认为的观点是 $\mu<400$，即产品说明书不真实。相反，就是原假设的内容。所以研究者的原假设与备择假设应为

H_0： $\mu\geqslant400$

H_1： $\mu<400$

【例 6.3】 某家企业的研究机构估计，该企业产品的市场占有比例超过 20%。为验证这一估计是否正确，该研究机构随机抽取了一个样本进行检验。试陈述用于检验的原假设与备择假设。

解：设该企业产品的市场占有比例真值为 π。显然，研究者就是这家企业的研究机构，他们认为的观点是“产品市场占有比例超过 20%”。因此，研究者建立的原假设和备择假设应为

H_0： $\pi\leqslant20\%$

H_1： $\pi>20\%$

下面把以上的 3 个例子汇总到表 6-2 格中，来总结一下原假设和备择假设的一些特点。

表 6-2 3 个例题的原假设和备择假设

例题	原假设和备择假设
例 6.1	H_0： μ=100； H_1： $\mu\neq100$
例 6.2	H_0： $\mu\geqslant400$； H_1： $\mu<400$
例 6.3	H_0： $\pi\leqslant20\%$； H_1： $\pi>20\%$

(1) 原假设和备择假设是一个完备事件组，而且相互对立。

从上面的 3 个例题可以看出，原假设和备择假设的集合区域是一个完备事件组，且是相互对立的，也就是说，在一项假设检验中，原假设和备择假设必有一个成立，而且只有一个成立。

(2) 在建立假设时，通常先确定备择假设，然后再确定原假设。

(3) 在假设检验中，“=”总是放在原假设上。

从上面的 3 个例子可以得知，“=”出现在原假设上，决不会出现在备择假设中。

(4) 假设检验的目的主要是收集证据拒绝原假设。

研究人员想证明的是他的观点，而他认为正确的观点作为备择假设，所以假设检验的目的主要是收集证据拒绝原假设。

2) 构造检验的统计量，并计算其值

在提出具体的假设之后，研究者需要提供可靠的证据来支持他所提出的备择假设。在实际操作中，提出证据的信息主要是来自所抽取的样本，假设检验就是要凭借可获得的样本观测结果帮助研究者做出最后的判断和决策。此时，有一个很自然的想法是，如果样本提供的证据能够证明原假设是不真实的，那么研究者就有理由拒绝它，并倾向于选择备择假设。既然研究者都倾向于通过样本信息对备择假设提供支持，并倾向于做出“拒绝原假设”的结论，那么研究此类问题，往往需要对这些信息进行压缩和提炼，即检验统计量便是对样本信息进行压缩和概括的结果。

定义 6.5　根据样本观测结果计算得到的，并据以对原假设和备择假设做出决策的某个样本统计量，称为检验统计量。

检验统计量实际上是总体参数的点估计量(如样本均值 $\bar{x}$ 就是总体均值 μ 的一个点估计量)，但点估计量并不能直接作为检验的统计量，只有将其标准化后，才能用于度量它与原假设的参数值之间的差异程度。

对点估计量进行标准化的依据有两个：①原假设 H_0 为真；②点估计量的抽样分布。

通常将标准化检验统计量简称为检验统计量，即检验统计量是服从标准正态分布、χ^2 分布、t 分布和 F 分布。

例如，对于总体均值和总体比例的检验，标准化检验统计量可表示为

$$\text{标准化检验统计量}=\frac{\text{点估计量}-\text{假设值}}{\text{点估计量的抽样标准差}}$$

检验统计量是一个随机变量，它的具体数值随着样本观测结果的不同而不同，但只要已知一组特定的样本观测结果，检验统计量的值也就唯一确定了。

检验统计量要求不能含有未知数，若含有，则无法计算出其值。

3) 根据给出的显著性水平 α，确定拒绝原假设 H_0 的区域

定义 6.6　能够拒绝原假设的检验统计量的所有可能取值的集合，称为拒绝域。

拒绝域就是由显著性水平 α 所围成的区域。如果利用样本观测结果计算出来的检验统计量的具体数值落在了拒绝域内，就拒绝原假设，否则就不拒绝原假设。

拒绝域的大小与事先选定的显著性水平有一定的关系。在确定了显著性水平 α 之后，就可以根据 α 值的大小确定出拒绝域的具体边界值。拒绝域的边界值称为临界值。在如何确定临界值之前，先介绍一下什么是显著性水平 α。

(1) 显著性水平α。

假设检验的目的是要根据样本信息做出决策，也就是做出是否拒绝原假设而倾向于备择假设的决策。显然，研究者希望做出正确的决策，但由于决策建立在样本信息的基础之上，而样本又是随机的，因而研究者就有可能犯错误。

如前所述，原假设与备择假设不能同时成立，即要么拒绝原假设H_0，要么不拒绝H_0。此时，研究人员希望的情况是，当原假设H_0正确时不拒绝它，当原假设H_0不正确时拒绝它。但是，很难保证不犯错误。假设检验过程中可能发生以下两类错误。

定义 6.7　当原假设正确时拒绝原假设，所犯的错误称为第Ⅰ类错误，又称弃真错误。犯第Ⅰ类错误的概率通常记为α。

定义 6.8　当原假设错误时不拒绝原假设，所犯的错误称为第Ⅱ类错误，又称取伪错误。犯第Ⅱ类错误的概率通常记为β。

假设检验中的结论及其后果有以下两种情况，见表 6-3 所示。

表 6-3　假设检验的结论与后果

决策结果	实际情况	
	H_0 正确	H_0 不正确
未拒绝 H_0	正确决策	第Ⅱ类错误 β
拒绝 H_0	第Ⅰ类错误 α	正确决策

需要注意的是，当样本容量一定时，不能同时减少α和β，即可以不犯第Ⅰ类错误或不犯第Ⅱ类错误，但难以保证两类错误都不犯。因为这两类错误的概率之间存在如下关系。

在样本容量不变的情况下，要减小α就会使β增大，而要增大α就会使β减小，这两类错误就像一个跷跷板。自然，人们希望犯两类错误的概率都尽可能小，但实际上很难做到。要使α和β同时减小的唯一办法是增加样本容量，但样本容量的增加又会受到许多因素的限制，所以人们只能在这两类错误发生的概率之间进行平衡，以使α与β控制在能够接受的范围内。

一般来说，对于一个给定的样本，如果犯第Ⅰ类错误的代价比犯第Ⅱ类错误的代价高，则将犯第Ⅰ类错误的概率定得低些较为合理；反之，如果犯第Ⅰ类错误的代价比犯第Ⅱ类错误的代价低，则将犯第Ⅰ类错误的概率定得高些。

至于假设检验中应先控制哪类错误，一般来说，发生哪一类错误的后果更为严重，就应该首先控制那类错误发生的概率。但是，由于犯第Ⅰ类错误的概率可由研究者进行控制，因此在假设检验中，人们往往先控制第Ⅰ类错误的发生概率。

发生第Ⅰ类错误的概率也常用于检验结论的可靠性度量，并将这一概率称为显著性水平。

定义 6.9　假设检验中发生第Ⅰ类错误的概率，称为显著性水平，记为α。

常用的显著性水平有α=0.01、α=0.05、α=0.1等，当然也可以取其他值。

(2) 确定临界值。

① 检验的方向。

在假设检验中，研究者感兴趣的备择假设的内容，可以是原假设 H_0 在某一特定方向的变化，也可以是一种没有特定方向的变化。例如，在例 6.2 中，研究者感兴趣的是奶粉的净含量是否低于 400g。同样，在例 6.3 中，研究者感兴趣的产品的市场占有比例是否超过 20%。这种具有方向性的假设称为单侧检验(或称单尾检验)。相反，在例 6.1 中，研究者感兴趣的备择假设没有特定的方向，他们只是关心备择假设 H_1 是否不同于原假设 H_0，而不关心 H_1 是大于还是小于 H_0，这种没有特定方向的假设称为双侧检验(或称双尾检验)。

定义 6.10　备择假设具有特定的方向性，并含有符号“＞”或“＜”的假设检验，称为单侧检验或单尾检验。

定义 6.11　备择假设没有特定的方向性，并含有符号“≠”的假设检验，称为双侧检验或双尾检验。

其中，在单侧检验中，由于研究者感兴趣的方向不同，又可分为左侧检验和右侧检验。如果研究者感兴趣的备择假设的方向为“＜”，称为左侧检验；如果研究者感兴趣的备择假设的方向为“＞”，称为右侧检验。

例如，设 μ 为总体参数(这里代表总体均值)，μ_0 为假设的参数的具体数值，则假设检验的基本形式总结见表 6-4 所示。

表 6-4　假设检验的基本形式

假设	双侧检验	单侧检验	
		左侧检验	右侧检验
原假设	$H_0:\ \mu=\mu_0$	$H_0:\ \mu\geqslant\mu_0$	$H_0:\ \mu\leqslant\mu_0$
备择假设	$H_1:\ \mu\neq\mu_0$	$H_1:\ \mu<\mu_0$	$H_1:\ \mu>\mu_0$

② 确定临界值。

定义 6.12　根据给定的显著性水平确定的拒绝域的边界值，称为临界值。

在给定显著性水平 α 和检验统计量的分布后，查一些常用统计表就可以得到具体的临界值或利用 Excel 中的统计函数也可以得出此临界值。

如果双侧检验的拒绝域在抽样分布的两侧(所以称为双侧检验)。在单侧检验中，如果备择假设具有符号“＜”，那么拒绝域位于抽样分布的左侧，称为左侧检验；如果备择假设具有符号“＞”，那么拒绝域位于抽样分布的右侧，称为右侧检验。

在给定显著性水平 α 的条件下，拒绝域和临界值可用图 6.1 来表示。

4) 统计决策

由图 6.1 可以得出利用统计量进行检验时的准则。

(1) 双侧检验：|统计量|＞临界值，拒绝原假设。

(2) 左侧检验：统计量的值＜-临界值，拒绝原假设。

(3) 右侧检验：统计量的值＞临界值，拒绝原假设。

注意：在假设检验中，应对原假设 H_0 采取“拒绝”或“不拒绝”的表述方式，而不采取“接受”的表述方式。“不拒绝”的表述实际上意味着并未给出明确的结论，原假设正

确与否尚未确定。如果说“接受”原假设，则意味着已经证明了原假设是正确的；但实际上，假设检验并不提供原假设“正确”的证据，它只提供不利于原假设的证据。

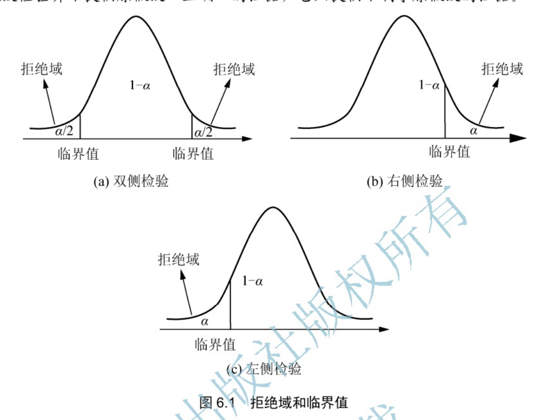


图 6.1 拒绝域和临界值

2. 利用 P 值法进行决策

1) 利用 P 值进行决策的步骤

(1) 提出原假设 H_0 和备择假设 H_1。

(2) 构造检验的统计量，并计算其值。

(3) 根据检验统计量的值计算 P 值。

(4) 统计决策。

因为此步骤与传统的假设检验的前两个步骤相同，这里不再重复介绍，只介绍第三步和第四步的内容。

2) P 值的计算

传统的统计量检验方法是在检验之前确定显著性水平 α，这就意味着事先确定了拒绝域。这样一来，不论检验统计量的值是大还是小，只要它的值落入拒绝域就拒绝原假设 H_0，否则就不拒绝原假设 H_0。这种固定的显著性水平 α 对检验结果的可靠性起一种度量作用。但不足的是，α 是犯第 I 类错误的上限控制值，它只能提供检验结论可靠性的一个大致范围；但对于一个特定的假设检验问题，它却无法给出观测数据与原假设之间不一致程度的精确度量。也就是说，仅从显著性水平来比较，如果选择的 α 值相同，那么所有检验结论的可靠性都一样。要测量出样本观测数据与原假设中所假设的值 μ_0 的偏离程度，就需要计算 P 值。

定义 6.13 如果原假设 H_0 是正确的，那么所有的样本结果出现实际观测结果那么极端的概率，称为 P 值，也称观察到的显著性水平。

下面来看 P 值的计算过程。

为理解 P 值的计算过程，统一使用符号 Z 表示检验统计量，Z_c 表示根据样本数据计算得到的检验统计量值，对于假设检验的 3 种基本形式，从抽样分布上看，计算 P 值的一般表达式如下。

(1) 左侧检验。

$$H_0:\ \mu \geqslant \mu_0;\ H_1:\ \mu < \mu_0$$

当 $\mu=\mu_0$，P 值的计算公式为

$$P = P(Z \leqslant Z_c | \mu=\mu_0) \tag{6.1}$$

(2) 右侧检验。

$$H_0:\ \mu \leqslant \mu_0;\ H_1:\ \mu > \mu_0$$

当 $\mu=\mu_0$，P 值的计算公式为

$$P = P(Z \geqslant Z_c | \mu=\mu_0) \tag{6.2}$$

(3) 双侧检验。

$$H_0:\ \mu=\mu_0;\ H_1:\ \mu \neq \mu_0$$

当 $\mu=\mu_0$ 时，P 值的计算公式为：

$$P = 2P(Z \geqslant |Z_c| | \mu=\mu_0) \tag{6.3}$$

为了理解不同检验的 P 值计算，可以用图 6.2 来表示。

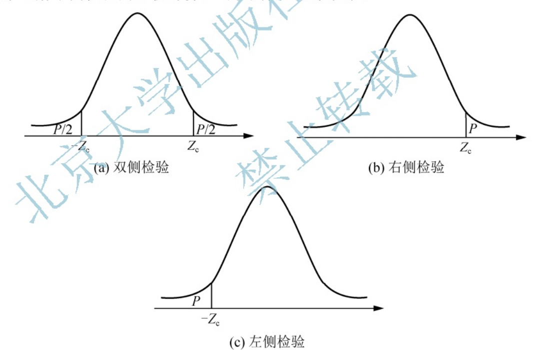


图 6.2　不同检验的 P 值

P 值的计算在计算机软件计算十分容易。

3) 利用 P 值决策的规则

利用 P 值进行决策的规则十分简单。不论是单侧检验还是双侧检验，使用 P 值进行决策的准则都是：如果 $P<\alpha$，拒绝 H_0；如果 $P>\alpha$，不拒绝 H_0。但在现代统计检验中，如果 $P<0.1$ 代表有“一些证据”不利于原假设；如果 $P<0.05$ 代表有“适度证据”不利于原假设；如果 $P<0.01$ 代表有“很强证据”不利于原假设。不再严格的给出显著性水平与 P 值进行比较。

6.2　一个总体参数的假设检验

本节将在 6.1 节的基础上介绍假设检验的具体应用，对于一个总体参数的假设检验包括总体均值 μ 、总体比例 π 和总体方差 σ^2 。

6.1 节介绍的所有概念都适用于下面介绍的检验方法，对于检验的步骤中，只有第 2 步，由于检验的参数不同，因此计算检验统计量的方法有所不同。本节的所有例题都采用了两种方法进行统计决策。

6.2.1　一个总体均值的假设检验

一个总体均值的假设检验要区分总体是否服从正态分布、总体方差 σ^2 是否已知等几种情况。

1. 正态总体、方差已知，或非正态总体、大样本

当总体服从正态分布且 σ 已知，或者总体不是正态分布但为大样本时，样本均值 $\bar{x}$ 的抽样分布均为正态分布，其数学期望为总体均值 μ ，方差为 $\frac{1}{n}\sigma^2$ ，即 $\bar{x} \sim N(\mu,\frac{1}{n}\sigma^2)$ ，所以采用正态分布的检验统计量。设假设的总体均值 μ_0 ，可以证明，样本均值经过标准化后服从标准正态分布，当总体方差 σ^2 已知时，总体均值的统计量为

$$Z=\frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}} \sim N(0,1) \tag{6.4}$$

如果是双侧检验，则拒绝域为 $(-\infty,-Z_{\alpha/2})\cup(Z_{\alpha/2},+\infty)$ ；如果是左侧单侧检验，则拒绝域为 $(-\infty,-Z_{\alpha})$ ；如果是右侧单侧检验，则拒绝域为 $(Z_{\alpha},+\infty)$ 。其中临界值可以查相应的统计分布表或利用 Excel 的统计函数计算出来。

【例 6.4】 某种袋装食品采用自动生产线生产，每袋的重量是 255g，标准差为 5g。为检验每袋重量是否符合要求，质检人员在某天生产的食品中随机抽取 40 袋进行检验，测得每袋平均重量为 255.8g。取显著性水平为 α=0.05，检验该天生产的食品是否符合标准要求。

解：

第一种方法：传统的假设检验，步骤如下。

(1) 提出的原假设和备择假设为

H_0： μ=255； H_1： $\mu \neq 255$

(2) 构造检验统计量，并计算其值。

$$Z=\frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}}=\frac{255.8-255}{5/\sqrt{40}}=1.01$$

(3) 根据给定的显著性水平 α=0.05，查标准正态分布表可知：

$Z_{\alpha/2} = Z_{0.025} = 1.96$

所以拒绝域为 $(-\infty,-1.96)\cup(1.96,+\infty)$ ，或者临界值利用 Excel 计算，操作的过程如下。

第一步：进入 Excel 表格界面，单击“插入函数”按钮，弹出“插入函数”对话框，在对话框中单击“或选择类别”的下拉按钮，在弹出的下拉列表中选择“统计”选项，并

在“选择函参数”列表中选择 NORM.S.INV 选项，单击“确定”按钮，弹出“函数参数”对话框。

第二步：在“函数参数”对话框中 Probability 文本框中输入“0.975”，得到函数值“1.959963985”，如图 6.3 所示，保留两位小数取 1.96。

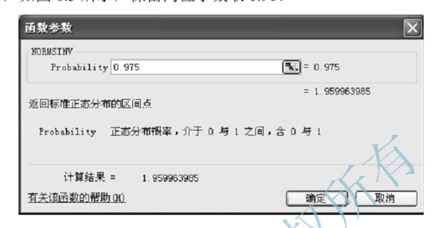


图 6.3 “函数参数”对话框 1

(4) 统计决策。

由于 $-1.96 < 1.01 < 1.96$，因此不拒绝原假设。

检验结果表明：样本提供的证据还不足以推翻原假设，因此不能证明该天生产的食品不符合标准要求。

第二种方法：利用 P 值进行统计决策

P 值的计算利用 Excel 中的统计函数功能计算，具体操作的步骤如下。

第一步：进入 Excel 表格界面，单击“插入函数”按钮，弹出“插入函数”对话框，在对话框中单击“或选择类别”的下拉按钮，在弹出的下拉列表中选择“统计”选项，并在“选择函数”列表中选择 NORM.S.DIST 选项，单击“确定”按钮，弹出“函数参数”对话框。

第二步：在“函数参数”对话框中的 Z 文本框中输入“1.01”，得到函数值“0.843752355”，如图 6.4 所示。



图 6.4 “函数参数”对话框 2

函数值“0.843752355，表示在标准正态分布条件下值为 1.01 左边的面积。

由于是双侧检验，因此有 $P = 2\times(1-0.843\ 752\ 355) = 0.312\ 495$ 。

统计决策：$P=0.312\,495>\alpha=0.05$，所以不拒绝原假设，和上一种方法的结论相同。

2. 正态总体、方差未知，或非正态总体，但样本是大样本

如果总体服从正态分布且σ未知，或总体并不服从正态分布，但只要是在大样本条件下，样本均值$\bar{x}$同样服从正态分布，即$\bar{x}\sim N(\mu,\frac{1}{n}\sigma^2)$。设假设的总体均值$\mu_0$，可以证明，样本均值经过标准化后服从标准正态分布，经过标准化以后的随机变量还是服从标准正态分布，即

$$Z=\frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}}\sim N(0,1)$$

但此时的标准正态分布的统计量中包含了一个未知参数(σ)，所以此时的统计量无法作为检验的统计量，但由于样本是大样本，前面学习过一致性，随着样本容量的不断增加，样本所计算出的样本统计量非常接近于总体参数，所以可以用样本方差s^2代替总体方差σ^2，即当总体方差未知，但为大样本时，此时总体均值检验的统计量为

$$Z=\frac{\bar{x}-\mu_0}{s/\sqrt{n}} \tag{6.5}$$

如果是双侧检验，则拒绝域为$(-\infty,-Z_{\alpha/2})\cup(Z_{\alpha/2},+\infty)$；如果是左侧单侧检验，则拒绝域为$(-\infty,-Z_{\alpha})$；如果是右侧单侧检验，则拒绝域为$(Z_{\alpha},+\infty)$。

【例 6.5】 一种机床加工的零件尺寸绝对平均误差为 1.35mm。生产厂家采用一种新的机床进行加工以期待进一步减低误差。为检验新机床加工的零件平均误差与旧机床相比是否显著降低，从某天生产的零件中随机抽取 50 个进行检验，其绝对误差的平均数为$\bar{x}=1.215\,2\text{mm}$，标准差$s=0.365\,749\text{mm}$，试检验新机床加工的零件尺寸的平均误差与旧机床相比是否显著降低。($\alpha=0.05$)

解：

第一种方法：传统的假设检验，步骤如下。

(1) 提出原假设和备择假设。

H_0：$\mu\geqslant1.35$；H_1：$\mu<1.35$

(2) 构造检验统计量，并计算其值。

$$Z=\frac{\bar{x}-\mu_0}{s/\sqrt{n}}$$

其中根据样本数据计算得

$$\bar{x}=1.215\,2,\quad s=0.365\,749$$

计算检验统计量的具体数值为

$$Z=\frac{\bar{x}-\mu_0}{s/\sqrt{n}}=\frac{1.215\,2-1.35}{0.365\,749/\sqrt{50}}=-2.606\,1$$

(3) 根据给定的显著性水平α=0.01，查标准正态分布表或利用 Excel 的统计函数(与例 6.4 操作相同)可知：

$$Z_{\alpha}=Z_{0.01}=2.33$$

所以拒绝域为$(-\infty,-2.33)$。

(4) 统计决策。

$$Z=-2.6061< -Z_{0.01}=-2.33$$

所以拒绝原假设。该检验结构表明，新机床加工的零件尺寸的平均误差与旧机床相比有显著降低。

第二种方法：利用 P 值进行决策(操作过程同例 6.4)。

计算出 $P=0.004\,578\,986$，所以统计决策 $P=0.004\,578\,986<\alpha=0.01$，所以拒绝原假设。该结论与统计量检验一致。

【例 6.6】 某玉米品种的平均产量为 5 100kg/hm^2。一家研究机构对玉米品种进行了改良后以期待提高产量。为检验改良后的玉米产量是否有显著提高，随机抽取了 49 个地块进行试种，得到的样本平均产量为 5 275kg/hm^2，标准差为 140 kg/hm^2。试检验改良后的玉米产量是否有显著提高。(α=0.05)

解：

第一种方法：传统的假设检验，步骤如下。

(1) 提出原假设和备择假设。

$$H_0:\ \mu\leqslant 5100;\ H_1:\ \mu>5100$$

(2) 构造检验统计量，并计算其值。

$$Z=\frac{\bar{x}-\mu_0}{s/\sqrt{n}}=\frac{5\,275-5\,100}{140/\sqrt{49}}=8.75$$

(3) 根据给定的显著性水平 α=0.05，查标准正态分布表或利用 Excel 的统计函数计算(与例 6.4 操作相同)可知：

$$Z_\alpha=Z_{0.05}=1.645$$

所以拒绝域为 $(1.645,+\infty)$。

(4) 统计决策

由于 $Z=8.75>Z_{0.05}=1.645$，因此拒绝原假设。检验结果表明，改良后的玉米产量有显著提高。同样利用 Excel 计算出来的 P 值为 0.000088＜α=0.05，同样拒绝原假设。

大样本情况下一个总体均值的检验方法汇总见表 6-5 所示。

表 6-5　大样本情况下一个总体均值的检验方法

项目	双侧检验	左侧检验	右侧检验
假设形式	$H_0:\ \mu=\mu_0$； $H_1:\ \mu\neq\mu_0$	$H_0:\ \mu\geqslant\mu_0$； $H_1:\ \mu<\mu_0$	$H_0:\ \mu\leqslant\mu_0$； $H_1:\ \mu>\mu_0$
检验统计量	σ 已知时：$Z=\frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}}$；$\sigma$ 未知时：$Z=\frac{\bar{x}-\mu_0}{s/\sqrt{n}}$		
α 与拒绝域	$(-\infty,-Z_{\alpha/2})\cup(Z_{\alpha/2},+\infty)$	$(-\infty,-Z_\alpha)$	$(Z_\alpha,+\infty)$
P 值决策准则	$P<\alpha$，拒绝原假设		

3. 正态总体、方差未知，样本是小样本

在小样本($n<30$)情形下，检验统计量的选择与总体是否服从正态分布、总体方差是否已知有着密切联系。

设假设的总体均值为μ_0，总体服从正态分布，无论大小样本，可以证明，样本均值经过标准化后服从标准正态分布，经过标准化以后的随机变量还是服从标准正态分布，即

$$Z=\frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}}\sim N(0,1)$$

此时该统计量中包含了一个未知参数(σ)，所以无法作为μ的检验统计量。又因为是小样本，不能直接用样本方差s^2代替σ^2，此时从标准正态无法得出作为总体μ的检验统计量。若用χ^2分布，χ^2是标准正态分布的平方加和，经过平方加和后，其式中还有一个未知参数(σ)，还是不可以。此时可用t分布。推导过程如下：

$$\chi^2=\frac{(n-1)s^2}{\sigma^2}\sim\chi^2(n-1)$$

所以有

$$t=\frac{Z}{\sqrt{\chi^2/(n-1)}}\sim t(n-1)$$

把上两式代入得到

$$t=\frac{\dfrac{\bar{x}-\mu_0}{\sigma/\sqrt{n}}}{\sqrt{\dfrac{(n-1)s^2}{\sigma^2}/(n-1)}}\sim t(n-1)$$

整理后得到当总体方差未知，但为大样本时，此时总体均值检验的统计量为

$$t=\frac{\bar{x}-\mu_0}{s/\sqrt{n}}\sim t(n-1) \tag{6.6}$$

如果是双侧检验，则拒绝域为$(-\infty,-t_{\frac{\alpha}{2}}(n-1))\cup(t_{\frac{\alpha}{2}}(n-1),+\infty)$；如果是左侧单侧检验，则拒绝域为$(-\infty,-t_{\alpha}(n-1))$；如果是右侧单侧检验，则拒绝域为$(t_{\alpha}(n-1),+\infty)$。其中临界值可以查$t$分布表或利用 Excel 中的统计函数计算。

小样本情况下一个总体均值的检验方法汇总见表 6-6 所示。

表 6-6　小样本情况下一个总体均值的检验方法

项目	双侧检验	左侧检验	右侧检验
假设形式	H_0：$\mu=\mu_0$； H_1：$\mu\neq\mu_0$	H_0：$\mu\geqslant\mu_0$； H_1：$\mu<\mu_0$	H_0：$\mu\leqslant\mu_0$； H_1：$\mu>\mu_0$
检验统计量	σ 已知时：$t=\frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}}$；$\sigma$ 未知时：$t=\frac{\bar{x}-\mu_0}{s/\sqrt{n}}$		
α 与拒绝域	$(-\infty,-t_{\alpha/2}(n-1))\cup(t_{\alpha/2}(n-1),+\infty)$	$(-\infty,-t_{\alpha}(n-1))$	$(t_{\alpha}(n-1),+\infty)$
P 值决策准则	$P<\alpha$，拒绝原假设		

【例 6.7】 一种机床加工的零件尺寸平均长度要求为 12cm，高于或低于该标准均被认为是不合格的。购买该零件的企业在购进零件时，通常是经过招标，然后对中标的零件提供商的样品进行检验，以决定是否采购。某汽车生产企业对一个零件提供商提供的 12 个样本进行了检验，其结果见表 6-7 所示。

表 6-7 某汽车生产企业的样本零件的长度数据

12.2	10.8	12.0	11.8	11.9	12.4	11.3	12.2	12.0	12.3	11.9	12.4

假定该供货商的零件长度服从正态分布，那么在α=0.05的显著性水平下，检验该供货商提供的零件是否符合要求。

解：

第一种方法：传统的假设检验，步骤如下。

(1) 提出原假设和备择假设。

H_0： μ=12； H_1： $\mu \neq 12$

(2) 构造检验统计量，并计算其值。

$$t=\frac{\bar{x}-\mu_0}{s/\sqrt{n}}$$

其中根据样本数据计算得：$\bar{x}=11.9$， $s=0.469\,687$，计算检验统计量的具体数值为

$$t=\frac{\bar{x}-\mu_0}{s/\sqrt{n}}=\frac{11.9-12}{0.469\,687/\sqrt{12}}=-0.737\,5$$

(3) 显著性水平$\alpha=0.05$，根据自由度$n-1=12-1=11$，查t分布表可知：

$$t_{\alpha/2}(n-1)=t_{0.025}(11)=2.593$$

所以拒绝原假设的区域为$(-\infty,-2.593)\cup(2.593,+\infty)$；或者利用 Excel 的统计函数计算可得，操作过程如下。

第一步：进入 Excel 表格界面，单击“插入函数”按钮，弹出“插入函数”对话框，在对话框中单击“或选择类别”的下拉按钮，在弹出的下拉列表中选择“统计”选项，并在“选择函参数”列表中选择 T.INV 选项，单击“确定”按钮，弹出“函数参数”对话框。

第二步：在“函数参数”对话框中的 Probability 文本框中输入“0.025”，在 Deg_freedom 文本框中输入“11”，得到函数值“2.593092681”，如图 6.5 所示，保留 3 位小数取 2.593。



图 6.5 “函数参数”对话框 3

(4) 统计决策。

由于$|t|=0.737\,5 < t_{0.025}(11)=2.593$，因此不拒绝原假设，即样本提供的证据还不足以

推翻原假设。

第二种方法：利用 P 值进行决策，其操作过程如下。

第一步：进入 Excel 表格界面，单击“插入函数”按钮，弹出“插入函数”对话框，在对话框中单击“或选择类别”的下拉按钮，在弹出的下拉列表中选择“统计”选项，并在“选择函数”列表中选择 T.DIST 选项，单击“确定”按钮，弹出“函数参数”对话框。

第二步：在“函数参数”对话框中的 X 文本框中输入“0.7375”，在 Deg_freedom 文本框中输入“11”，在“Tails”文本框中输入“2”，可得出统计值“0.476256199”如图 6.6 所示。

由于 P 值 0.476 256 199＞0.05，因此不拒绝原假设。

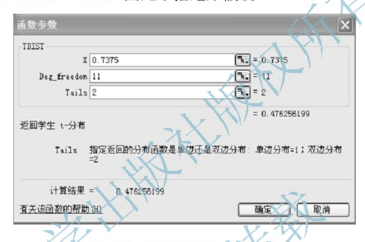


图 6.6 “函数参数”对话框 4

6.2.2 一个总体比例的假设检验

这里只考虑大样本情形下的总体比例检验。所以总体比例检验的 3 种基本形式如下。

(1) 双侧检验：H_0：$\pi=\pi_0$；H_1：$\pi \neq \pi_0$。

(2) 左侧检验：H_0：$\pi \geqslant \pi_0$；H_1：$\pi < \pi_0$。

(3) 右侧检验：H_0：$\pi \leqslant \pi_0$；H_1：$\pi > \pi_0$。

在构造检验统计量时，仍然利用样本比例 p 与总体比例 π 之间的距离等于多少个标准差 σ_p 来衡量。这是因为，在大样本情形下，统计量 p 近似服从正态分布，即设总体比例 π 的假设值 π_0，检验的统计量为

$$Z=\frac{p-\pi_0}{\sqrt{\dfrac{\pi_0(1-\pi_0)}{n}}} \sim N(0,1) \tag{6.7}$$

大样本情况下总体比例检验的方法见表 6-8 所示。

表 6-8　大样本情况下总体比例检验的方法

项目	双侧检验	左侧检验	右侧检验
假设形式	H_0： $\pi=\pi_0$； H_1： $\pi\neq\pi_0$	H_0： $\pi\geqslant\pi_0$ H_1： $\pi<\pi_0$	H_0： $\pi\leqslant\pi_0$； H_1： $\pi>\pi_0$
检验统计量	$Z=\dfrac{p-\pi_0}{\sqrt{\dfrac{\pi_0(1-\pi_0)}{n}}}$		
α 与拒绝域	$(-\infty,-Z_{\alpha/2})\cup(Z_{\alpha/2},+\infty)$	$(-\infty,-Z_{\alpha})$	$(Z_{\alpha},+\infty)$
P 值决策准则	$P<\alpha$，拒绝原假设		

【例 6.8】 一家研究机构，声称该城市拥有汽车比例超过 40%。为验证这一说法是否属实，这家研究机构部门抽取了由 200 个家庭组成的一个随机样本，发现有 87 个家庭拥有汽车。取显著性水平 $\alpha=0.05$，检验该城市拥有汽车比例是否超过 40%。

解：

第一种方法：传统的假设检验，步骤如下。

(1) 提出的原假设和备择假设为

H_0： $\pi\leqslant 40\%$； H_1： $\pi>40\%$

(2) 依据题意，大样本，所以检验的统计量为

$$Z=\frac{p-\pi_0}{\sqrt{\dfrac{\pi_0(1-\pi_0)}{n}}}$$

根据样本结果计算得 $P=\dfrac{87}{200}=43.5\%$，故检验统计量为

$$Z=\frac{0.435-0.4}{\sqrt{\dfrac{0.4\times(1-0.4)}{200}}}=\frac{0.035}{0.035}=1$$

(3) 根据显著性水平 $\alpha=0.05$，查标准正态分布表或利用 Excel 的统计函数计算(与例 6.4 操作相同)可得：

$$Z_{\alpha/2}=Z_{0.025}=1.96$$

所以拒绝域为 $(-\infty,-1.96)\cup(1.96,+\infty)$。

(4) 统计决策

由于 $|Z|=1<Z_{\alpha/2}=1.96$，因此不拒绝原假设。在显著性水平 $\alpha=0.05$ 的条件下，样本提供的证据表明该机构的说法并不属实。

第二种方法：利用 P 值进行统计决策，其操作过程同例 6.4。

计算出 P 值为 $0.156>\alpha=0.05$，不拒绝原假设。

6.2.3　一个总体方差的假设检验

对于多数生产和生活领域而言，仅仅保证所观测到的样本均值维持在特定水平范围之内并不意味着整个过程的正常，方差的大小是否适度则是需要考虑的另一个重要因素。一个方差大的产品自然意味着其质量或性能不稳定。因此，总体方差 σ^2 的检验也是假设检验

的重要内容之一。

用σ_0^2表示假定的总体方差的某一个取值，总体方差检验的 3 种基本形式如下。

(1) 双侧检验。H_0：$\sigma^2=\sigma^2_0$；H_1：$\sigma^2\neq\sigma^2_0$。

(2) 左侧检验。H_0：$\sigma^2\geqslant\sigma^2_0$；$H_1$：$\sigma^2<\sigma^2_0$。

(3) 右侧检验。H_0：$\sigma^2\leqslant\sigma^2_0$；$H_1$：$\sigma^2>\sigma^2_0$。

对总体方差进行检验，检验统计量要从其样本方差入手，设总体方差的假设值为σ_0^2，由前面的介绍可知其检验统计量为

$$\chi^2=\frac{(n-1)s^2}{\sigma_0^2}\sim\chi^2(n-1) \tag{6.8}$$

一个总体方差检验的方法见表 6-9 所示。

表 6-9　一个总体方差检验的方法

项目	双侧检验	左侧检验	右侧检验
假设形式	H_0：$\sigma^2=\sigma_0^2$； H_1：$\sigma^2\neq\sigma_0^2$	H_0 $\sigma^2\geqslant\sigma_0^2$； H_1：$\sigma^2<\sigma_0^2$	H_0：$\sigma^2\leqslant\sigma_0^2$； H_1：$\sigma^2>\sigma_0^2$
检验统计量	$\chi^2=\dfrac{(n-1)s^2}{\sigma_0^2}$		
α 与拒绝域	$(0,\chi^2_{1-\alpha/2}(n-1))\cup(\chi^2_{\alpha/2}(n-1),+\infty)$	$(0,\chi^2_{1-\alpha}(n-1))$	$(\chi^2_{\alpha}(n-1),+\infty)$
P 值决策准则	$P<\alpha$，拒绝原假设		

【例 6.9】 一个制造商所生产的零件直径的方差本来是 0.001 56mm^2。后来为削减成本，就采用了一种费用较低的生产方法。从新方法制造的零件中随机抽取 200 个作样本，测得零件直径的方差为 0.002 11mm^2。在显著性水平$\alpha=0.05$下，检验新方法生产零件的方差是否比老方法大。

解：传统的假设检验步骤如下。

(1) 提出原假设和备择假设。

$H_0:\sigma^2\leqslant0.001\,56$；$H_1:\sigma^2>0.001\,56$

(2) 构造检验的统计量，并计算其值。

$$\chi^2=\frac{(n-1)s^2}{\sigma_0^2}\sim\chi^2(n-1)$$

其中$s^2=0.002\,11$；$\sigma_0^2=0.001\,56$，则有

$$\chi^2=\frac{(200-1)\times0.002\,11}{0.001\,56}=269.16$$

(3) $\alpha=0.05$，确定拒绝原假设的区域。

根据题意可得拒绝原假设的区域为$(\chi_\alpha{}^2(199),+\infty)$，其中查$\chi^2$分布表，$\chi_{0.05}{}^2(199)=232.911\,8$。

(4) 统计决策。

由于$\chi^2=269.16>\chi_{0.05}{}^2(199)=232.911\,8$，所以拒绝原假设，新方法比老方法的方差大。

6.3　两个总体参数的假设检验

本节的检验主要介绍传统的假设检验的程序，一般不再给出拒绝域的图示。同时两个总体参数的假设检验的 Excel 操作过程将在案例中进行详细的介绍。

6.3.1　两个总体均值之差的假设检验

两个总体参数的检验包括两个总体均值之差 $\mu_1-\mu_2$ 的检验、两个总体比例之差 $\pi_1-\pi_2$ 的检验和两个总体方差之比 σ_1^2/σ_2^2 的检验等。检验的程序与一个总体参数的检验类似，但统计量的计算要复杂一些。

根据样本获得方式的不同，两个总体均值的检验分为独立样本和配对样本两种情形，而且也有大样本与小样本之分。检验的统计量是以两个样本均值之差 $\bar{x}_1-\bar{x}_2$ 的抽样分布为基础构造出来的。对于大样本和小样本两种情形，由于两个样本均值之差经标准化的分布不同，检验的统计量也略有差异。

1. 独立样本，两个总体正态分布，或非正态分布，大样本，且两个总体方差已知

当两个总体是正态分布，或非正态分布，但大样本时，从两个总体各自抽取的样本均值均服从正态分布，即 $\bar{x}_1 \sim N(\mu_1,\frac{1}{n_1}\sigma_1^2)$；$\bar{x}_2 \sim N(\mu_2,\frac{1}{n_2}\sigma_2^2)$，构造 $\mu_1-\mu_2$ 的检验统计量要从样本抽样分布 $\bar{x}_1-\bar{x}_2$ 为基础，所以有

$$\bar{x}_1-\bar{x}_2 \sim N(\mu_1-\mu_2,\frac{1}{n_1}\sigma_1^2+\frac{1}{n_2}\sigma_2^2)$$

设两个总体均值之差的假设值为 $(\mu_1-\mu_2)_0$ 经标准化后，可检验统计量为

$$Z=\frac{(\bar{x}_1-\bar{x}_2)-(\mu_1-\mu_2)_0}{\sqrt{\frac{1}{n_1}\sigma_1^2+\frac{1}{n_2}\sigma_2^2}} \sim N(0,1) \tag{6.9}$$

【例 6.10】 某企业公司对男女职员的每天平均工资进行了调查，男职员总体的方差为 $\sigma_1^2=64$，女职员总体的方差为 $\sigma_2^2=42.25$。独立抽取了具有同类工作经验的男女，职员的两个随机样本，并记录下两个样本的均值、样本容量数据见表 6-10 所示。在显著性水平为 0.05 的条件下，能否认为男职员与女职员的每天平均工资存在显著性差异？

表 6-10　两样本的数据结果

男职员	女职员
$n_1=44$	$n_2=32$
$\bar{x}_1=75$	$\bar{x}_2=70$

解：设 μ_1=男职员的平均小时工资；μ_2=女职员的平均小时工资。

(1) 提出原假设和备择假设。

$$H_0:\mu_1-\mu_2=0；\quad H_1:\mu_1-\mu_2\neq 0$$

(2) 构造检验的统计量，并计算其值。

由于两个样本是独立，且方差已知，所以检验的统计量为

$$Z=\frac{(\bar{x}_1-\bar{x}_2)-(\mu_1-\mu_2)_0}{\sqrt{\frac{1}{n_1}\sigma_1^2+\frac{1}{n_2}\sigma_2^2}}\sim N(0,1)$$

$$=\frac{(75-70)-0}{\sqrt{\frac{1}{44}\times 64+\frac{1}{32}\times 42.25}}=3.002$$

(3) 根据显著性水平 0.05，拒绝原假设的区域为

$$(-\infty,-Z_{0.025})\cup(Z_{0.025},+\infty)$$

查标准正态分布表得 $Z_{0.025}=1.96$，所以拒绝域为 $(-\infty,-1.96)\cup(1.96,+\infty)$。

(4) 统计决策。

因为有 $Z=3.002>Z_{0.025}=1.96$，所以拒绝原假设，即认为男职员与女职员的每天平均工资存在显著差异。

2. 独立样本，大样本，且两个总体方差未知

在大样本情况下，两个样本均值之差 $\bar{x}_1-\bar{x}_2$ 的抽样分布近似的服从正态分布，即有

$$\bar{x}_1-\bar{x}_2\sim N(\mu_1-\mu_2,\frac{1}{n_1}\sigma_1^2+\frac{1}{n_2}\sigma_2^2)$$

设两个总体均值之差的假设值为 $(\mu_1-\mu_2)_0$，经标准化后可得

$$Z=\frac{(\bar{x}_1-\bar{x}_2)-(\mu_1-\mu_2)_0}{\sqrt{\frac{1}{n_1}\sigma_1^2+\frac{1}{n_2}\sigma_2^2}}\sim N(0,1)$$

由于统计量中含有未知数($\sigma_1{}^2$和$\sigma_2{}^2$)，但两个样本是大样本，可分别用样本方差 s_1^2和s_2^2 替代，此时检验的统计计为

$$Z=\frac{(\bar{x}_1-\bar{x}_2)-(\mu_1-\mu_2)_0}{\sqrt{\frac{1}{n_1}s_1^2+\frac{1}{n_2}s_2^2}}\sim N(0,1) \tag{6.10}$$

【例 6.11】 某研究机构对两种组装产品的方法每小时组装产品数量进行了调查，独立抽取了具有同类工作经验工人的两个随机样本，并记录下两组工人的每小时组装产品的平均数量、样本容量和样本方差数据，见表 6-11 所示。在显著性水平为 0.05 的条件下，能否认为两种组装方法平均小时组装数量是否存在显著性差异？

表 6-11　两样本的数据

样本 1	样本 2
$n_1=44$	$n_2=32$
$\bar{x}_1=75$	$\bar{x}_2=70$
$s_1^2=4$	$s_2^2=7$

解：传统的假设检验的步骤如下。

(1) 提出原假设和备择假设。

$$H_0:\mu_1-\mu_2=0;\quad H_1:\mu_1-\mu_2\neq 0$$

(2) 检验的统计量，并计算其值。

由于两个样本是独立，且方差已知，所以检验的统计量为

$$Z=\frac{(\bar{x}_1-\bar{x}_2)-(\mu_1-\mu_2)_0}{\sqrt{\frac{1}{n_1}s_1^2+\frac{1}{n_2}s_2^2}}\sim N(0,1)$$

$$=\frac{(75-70)-0}{\sqrt{\frac{1}{44}\times 4+\frac{1}{32}\times 7}}=8.985\,2$$

(3) 根据显著性水平 0.05，拒绝原假设的区域为

$$(-\infty,-Z_{0.025})\cup(Z_{0.025},+\infty)。$$

查标准正态分布表得 $Z_{0.025}=1.96$，所以拒绝域为 $(-\infty,-1.96)\cup(1.96,+\infty)$。

(4) 统计决策。

因为有 $Z=8.985\,2>Z_{0.025}=1.96$，所以拒绝原假设，即认为两种组装产品的方法存在显著差异。

3. 独立小样本的检验，且两个总体方差未知

当两个样本都为独立小样本时，需要假定两个总体都服从正态分布，检验时有两种情况。

1) 两个总体的方差未知但相等时

无论两个总体的方差是否已知，只要两个总体是正态分布，其各自的样本均值均服从正态分布，即 $\bar{x}_1\sim N(\mu_1,\frac{1}{n_1}\sigma^2)$；$\bar{x}_2\sim N(\mu_2,\frac{1}{n_2}\sigma^2)$，则样本抽样分布 $\bar{x}_1-\bar{x}_2$ 也服从正态分布，有

$$\bar{x}_1-\bar{x}_2\sim N(\mu_1-\mu_2,(\frac{1}{n_1}+\frac{1}{n_2})\sigma^2)$$

经标准化后，可检验的统计量为

$$Z=\frac{(\bar{x}_1-\bar{x}_2)-(\mu_1-\mu_2)}{\sqrt{(\frac{1}{n_1}+\frac{1}{n_2})\sigma^2}}\sim N(0,1)$$

由于统计量中含有未知数(σ^2)，两个样本是小样本，不可分别用样本方差替代。

由于两个总体是服从正态分布，所以有

$$\frac{(n_1-1)s_1^2}{\sigma^2}\sim\chi^2(n_1-1)\qquad \frac{(n_2-1)s_2^2}{\sigma^2}\sim\chi^2(n_2-1)$$

χ^2 具有可加性，即有

$$\chi^2=\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{\sigma^2}\sim\chi^2(n_1+n_2-2)$$

根据上两个统计量，可构造出以下统计量：

$$t=\frac{Z}{\sqrt{\chi^2/(n_1+n_2-2)}}\sim t(n_1+n_2-2)$$

经整理得

$$t=\frac{\dfrac{(\overline{x}_1-\overline{x}_2)-(\mu_1-\mu_2)}{\sqrt{\left(\dfrac{1}{n_1}+\dfrac{1}{n_2}\right)\sigma^2}}}{\sqrt{\dfrac{(n_1-1)s_1^2+(n_2-1)s_2^2}{\sigma^2}/(n_1+n_2-2)}}\sim t(n_1+n_2-2)$$

设两个总体均值之差的假设值为 $(\mu_1-\mu_2)_0$ 经标准化后，可得检验统计量为

$$t=\frac{(\overline{x}_1-\overline{x}_2)-(\mu_1-\mu_2)_0}{\sqrt{\left(\dfrac{1}{n_1}+\dfrac{1}{n_2}\right)\dfrac{(n_1-1)s_1^2+(n_2-1)s_2^2}{(n_1+n_2-2)}}}\sim t(n_1+n_2-2) \tag{6.11}$$

2) 两个总体的方差未知但不相等时

两个样本均值之差经标准化后不再服从自由度为 n_1+n_2-2 的 t 分布，而是近似服从自由度为 v 的 t 分布，这时检验统计量为

$$t=\frac{(\overline{x}_1-\overline{x}_2)-(\mu_1-\mu_2)_0}{\sqrt{\dfrac{1}{n_1}s_1^2+\dfrac{1}{n_2}s_2^2}}\sim t(v) \tag{6.12}$$

该统计量的自由度为 v，其计算公式为

$$v=\frac{\left(\dfrac{s_1^2}{n_1}+\dfrac{s_2^2}{n_2}\right)^2}{\dfrac{\left(\dfrac{s_1^2}{n_1}\right)^2}{n_1-1}+\dfrac{\left(\dfrac{s_2^2}{n_2}\right)^2}{n_1-1}} \tag{6.13}$$

其中，自由度一般为整数，需要上式进行四舍五入取整数。

【例 6.12】 用甲、乙两种方法同时加工某种同类型的零件，已知两种方法加工的零件直径分别服从正态分布 $N(\mu_1,\sigma_1^2)$ 和 $N(\mu_2,\sigma_2^2)$。为比较两台机床的加工精度有无显著差异，分别独立抽取了甲种方法加工的 8 个零件和乙种方法加工的 7 个零件，通过测量得到的直径数据(表 6-12)，在 $\alpha=0.05$ 的显著性水平下，检验两台种方法加工的零件是否一致：$\sigma_1{}^2=\sigma_2{}^2$； $\sigma_1{}^2\neq\sigma_2{}^2$。

表 6-12　两种方法加工零件的数据

单位：cm

方法	零件直径							
甲	10.5	9.8	9.7	10.4	10.1	10.0	9.0	9.9
乙	10.7	9.8	9.5	10.8	10.4	9.6	10.2	—

解：第一种情况：$\sigma_1^{\ 2}=\sigma_2^{\ 2}$。

(1) 提出原假设和备择假设。

$$H_0: \mu_1 - \mu_2 = 0\text{；}\quad H_1: \mu_1 - \mu_2 \neq 0$$

(2) 构造检验的统计量，并计算其值。

$$t = \frac{(\overline{x}_1 - \overline{x}_2) - (\mu_1 - \mu_2)_0}{\sqrt{\left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)\dfrac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{(n_1 + n_2 - 2)}}} \sim t(n_1 + n_2 - 2)$$

其中，根据样本可计算出 $\overline{x}_1 = 9.925$，$\overline{x}_2 = 10.143$；$s_1^2 = 0.2164$，$s_2^2 = 0.2729$。
代入检验的统计量中，有

$$t = \frac{9.925 - 10.143}{\sqrt{(1//8 + 1/7) \times 0.242\,5}} = -0.855$$

(3) $\alpha = 0.05$，所以拒绝原假设的区域为

$$(-\infty, -t_{0.025}(13)) \cup (t_{0.025}(13), +\infty)$$

其中临界值查 t 分布表，可得 $t_{0.025}(13) = 2.532\,638$，所以拒绝域为 $(-\infty, -2.532\,638) \cup (2.532\,638, +\infty)$。

(4) 统计决策。

$|t| = 0.855 < t_{0.025}(13) = 2.532\,638$，所以不拒绝原假设。

第二种情况 $\sigma_1^{\ 2} \neq \sigma_2^{\ 2}$。

(1) 提出原假设和备择假设。

$$H_0: \mu_1 - \mu_2 = 0\text{；}\quad H_1: \mu_1 - \mu_2 \neq 0$$

(2) 构造检验的统计量，并计算其值。

$$t = \frac{(\overline{x}_1 - \overline{x}_2) - (\mu_1 - \mu_2)_0}{\sqrt{\dfrac{1}{n_1}s_1^2 + \dfrac{1}{n_2}s_2^2}} \sim t(v)$$

其中

$$v = \frac{\left(\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}\right)^2}{\dfrac{\left(\dfrac{s_1^2}{n_1}\right)^2}{n_1 - 1} + \dfrac{\left(\dfrac{s_2^2}{n_2}\right)^2}{n_1 - 1}}$$

$$t = \frac{9.925 - 10.143}{\sqrt{\dfrac{1}{8} \times 0.216\,4 + \dfrac{1}{7} \times 0.272\,9}} = \frac{-0.218}{\sqrt{0.066\,035\,714}} = -0.848\,3$$

其中 $v = \dfrac{\left(\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}\right)^2}{\dfrac{\left(\dfrac{s_1^2}{n_1}\right)^2}{n_1 - 1} + \dfrac{\left(\dfrac{s_2^2}{n_2}\right)^2}{n_1 - 1}} = \dfrac{\left(\dfrac{0.216\,4}{8} + \dfrac{0.272\,9}{7}\right)^2}{\dfrac{\left(\dfrac{0.216\,4}{8}\right)^2}{8 - 1} + \dfrac{\left(\dfrac{0.272\,9}{7}\right)^2}{7 - 1}} = 12.184\,1$，所以自由度为 12。

(3) $\alpha = 0.05$，所以拒绝原假设的区域为

$$(-\infty, -t_{0.025}(12)) \cup (t_{0.025}(12), +\infty)$$

其中临界值查 t 分布表可得 $t_{0.025}(12) = 2.560\,033$，所以拒绝域为 $(-\infty, -2.560\,033) \cup (2.560\,033, +\infty)$。

(4) 统计决策。

$|t| = 0.848\,3 < t_{0.025}(12) = 2.560\,033$，所以不拒绝原假设。

4. 配对样本的检验

配对样本的检验需要假定两个总体配对差值构成的总体服从正态分布，而且配对差是由差值总体中随机抽取的。

(1) 对于大样本情形，配对差值经标准化后服从标准正态分布，因此设两个总体均值之差 $(\mu_1 - \mu_2)$ 的假设值为 $(\mu_1 - \mu_2)_0$，其检验的统计量为

$$Z = \frac{\bar{d} - (\mu_1 - \mu_2)_0}{S_d / \sqrt{n}} \sim N(0,1) \tag{6.14}$$

式中，$\bar{d}$ 为配对差值的平均数；S_d 为配对差值的标准差。

(2) 对于小样本情形，配对差值经标准化后服从自由度为 $n-1$ 的 t 分布。因此设两个总体均值之差 $(\mu_1 - \mu_2)$ 的假设值为 $(\mu_1 - \mu_2)_0$，其检验的统计量为

$$t = \frac{\bar{d} - (\mu_1 - \mu_2)_0}{S_d / \sqrt{n}} \sim t(n-1) \tag{6.15}$$

式中，$\bar{d}$ 为配对差值的平均数；S_d 为配对差值的标准差。

【例 6.13】 随机抽取一个由 8 名教师组成的样本，要求每名教师对两种看法进行评分(0～10 分)。评分的数据见表 6-13 所示。取显著性水平 $\alpha = 0.05$，检验 8 名教师对两种看法的评分是否存在显著差异。

表 6-13　两种看法评分等级的样本数据

教师编号		1	2	3	4	5	6	7	8
评分等级	看法 1	5	4	7	3	5	8	5	6
	看法 2	6	6	7	4	3	9	7	6

解：设 μ_1=教师对看法 1 的平均评分，μ_2=教师对看法 2 的平均评分。

(1) 依题意建立的原假设和备择假设为

$$H_0: \mu_1 - \mu_2 = 0;\quad H_1: \mu_1 - \mu_2 \neq 0$$

(2) 由于是小样本，因此检验的统计量为

$$t = \frac{\bar{d} - (\mu_1 - \mu_2)_0}{S_d / \sqrt{n}} \sim t(n-1)$$

样本配对差值见表 6-14 所示。

表 6-14 样本配对差值

序号	看法 1	看法 2	差值
1	5	6	−1
2	4	6	−2
3	7	7	0
4	3	4	−1
5	5	3	2
6	8	9	−1
7	5	7	−2
8	6	6	0

根据样本可计算出 $\bar{d}=-0.625$；$S_d=1.302\,47$，代入检验统计量中，得

$$t=\frac{-0.625-0}{1.302\,47/\sqrt{8}}=\frac{-0.625}{0.460\,493}=-1.357\,24$$

(3) 根据给出的显著性水平 $\alpha=0.05$，则拒绝原假设的区域为

$$(-\infty,-t_{0.025}(7))\cup(t_{0.025}(7),+\infty)。$$

其中临界值查 t 分布表可得 $t_{0.025}(7)=2.841\,244$，所以拒绝域为 $(-\infty,-2.841\,244)\cup(2.841\,244,+\infty)$。

(4) 统计决策。

$-2.841\,244<-1.357\,24<2.841\,244$，所以不拒绝原假设。

6.3.2 两个总体比例之差的假设检验

两个总体比例之差$(\pi_1-\pi_2)$的检验思路与一个总体比例的检验类似，要求两个样本都是大样本。根据两个样本比例之差的抽样分布 $p_1\sim N\left(\pi_1,\frac{\pi_1(1-\pi_1)}{n_1}\right)$ 和 $p_2\sim N\left(\pi_2,\frac{\pi_2(1-\pi_2)}{n_2}\right)$，可以得到两个总体比例之差的检验的统计量为

$$Z=\frac{(p_1-p_2)-(\pi_1-\pi_2)}{\sqrt{\frac{\pi_1(1-\pi_1)}{n_1}+\frac{\pi_2(1-\pi_2)}{n_2}}}\sim N(0,1)$$

式中，$\sqrt{\frac{\pi_1(1-\pi_1)}{n_1}+\frac{\pi_2(1-\pi_2)}{n_2}}$ 是两个样本比例之差的标准误差，且含有未知数 π_1和π_2。需要利用两个样本比例 P_1和P_2 来估计。具体可分为两种情况。

1. 检验两个总体比例之差是否相等

检验两个总体比例之差是否相等，即原假设和备择假设的内容有 3 种。

(1) $H_0:\pi_1-\pi_2=0$；$H_1:\pi_1-\pi_2\neq 0$。

(2) $H_0:\pi_1-\pi_2\leqslant 0$；$H_1:\pi_1-\pi_2>0$。

(3) $H_0:\pi_1-\pi_2\geqslant 0$；$H_1:\pi_1-\pi_2<0$。

这时，$\sqrt{\frac{\pi_1(1-\pi_1)}{n_1}+\frac{\pi_2(1-\pi_2)}{n_2}}$ 最佳的估计量是将两个样本合并后得到的合并比例 P 替代 π_1 和 π_2。其中合并后的 P 计算过程如下。

设 x_1 表示样本 1 中具有某种属性的个体数，x_2 表示样本 2 中具有某种属性的个体数，则合并后的比例为

$$p=\frac{x_1+x_2}{n_1+n_2} \tag{6.16}$$

此时设两个总体比例之差的假设值为 $(\pi_1-\pi_2)_0$，检验的统计量为

$$Z=\frac{(p_1-p_2)-(\pi_1-\pi_2)_0}{\sqrt{\left(\frac{1}{n_1}+\frac{1}{n_2}\right)p(1-p)}} \sim N(0,1) \tag{6.17}$$

2. 检验两个总体比例之差等于某个常数

检验两个总体比例之差等于某个常数，即原假设和备择假设的内容有 3 种。

(1) $H_0:\pi_1-\pi_2=c$；$H_1:\pi_1-\pi_2\neq c$。

(2) $H_0:\pi_1-\pi_2\leqslant c$；$H_1:\pi_1-\pi_2>c$。

(3) $H_0:\pi_1-\pi_2\geqslant c$；$H_1:\pi_1-\pi_2<c$。

这时 $\sqrt{\frac{\pi_1(1-\pi_1)}{n_1}+\frac{\pi_2(1-\pi_2)}{n_2}}$ 中的 π_1 和 π_2，分别用各自的样本比例 p_1 和 p_2 来替代。此时设两个总体比例之差的假设值 $(\pi_1-\pi_2)_0$，检验的统计量为

$$Z=\frac{(p_1-p_2)-(\pi_1-\pi_2)_0}{\sqrt{\frac{p_1(1-p_1)}{n_1}+\frac{p_2(1-p_2)}{n_2}}} \sim N(0,1) \tag{6.18}$$

【例 6.14】 为了研究企业经理是否认为他们获得了成功，在随机抽取的 200 个企业的女性经理中，认为自己成功的人数为 48 人；而在对 95 个男经理的调查中，认为自己成功的人数为 39 人。在 $\alpha=0.05$ 的显著性水平下，检验男女经理认为自己成功的人数比例是否有显著差异。

解：设 π_1=女经理认为自己成功的比例；π_2=男经理认为自己成功的比例。

(1) 提出原假设和备择假设。

$$H_0:\pi_1-\pi_2=0；H_1:\pi_1-\pi_2\neq 0$$

(2) 构造检验的统计量，并计算其值。

根据题意，是检验两个总体比例之差是否相等的，所以检验的统计量为

$$Z=\frac{(p_1-p_2)-(\pi_1-\pi_2)_0}{\sqrt{p(1-p)\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}$$

其中，根据样本的数据，有 $p_1=48/200=0.24$；$p_2=39/95=0.41$；$p=\frac{48+39}{200+95}=0.295$，所以有

$$Z=\frac{0.24-0.41}{\sqrt{0.295\times(1-0.295)\left(\frac{1}{200}+\frac{1}{95}\right)}}=\frac{-0.17}{\sqrt{0.003\,229}}=\frac{-0.17}{0.056\,825}=-2.991\,639$$

(3) $\alpha=0.05$，确定拒绝原假设的区域为

$$(-\infty,-Z_{\alpha/2})\cup(Z_{\alpha/2},+\infty)$$

其中查标准正态分布表得，$Z_{0.025}=1.96$，所以拒绝域为$(-\infty,-1.96)\cup(1.96,+\infty)$

(4) 统计决策。

$|Z|=2.991\,639>Z_{0.025}=1.96$，所以拒绝原假设，即男、女经理认为自己成功的人数比例有显著差异。

【例 6.15】 承上例，在$\alpha=0.05$的显著性水平下，检验男经理比女经理认为自己成功的人数比例是否高于 15%。

解：设π_1=女经理认为自己成功的比例；π_2=男经理认为自己成功的比例。

(1) 提出原假设和备择假设。

$$H_0:\pi_2-\pi_1\leqslant 0.15;\ H_1:\pi_2-\pi_1>0.15$$

(2) 构造检验的统计量，并计算其值。

根据题意，检验两个总体比例之差是否等于一个常数，所以检验的统计量为

$$Z=\frac{(p_2-p_1)-(\pi_2-\pi_1)_0}{\sqrt{\frac{p_1(1-p_1)}{n_1}+\frac{p_2(1-p_2)}{n_2}}}$$

其中，根据样本的数据有$p_1=48/200=0.24$；$p_2=39/95=0.41$，所以有

$$Z=\frac{(0.41-0.24)-0.15}{\sqrt{\frac{0.24\times(1-0.24)}{200}+\frac{0.41\times(1-0.41)}{95}}}=\frac{0.02}{\sqrt{0.003\,458}}=\frac{0.02}{0.058\,807}=0.340\,093$$

(3) $\alpha=0.05$，确定拒绝原假设的区域为

$$(Z_\alpha,+\infty)$$

其中临界值查标准正态分布表得$Z_{0.05}=1.645$，所以拒绝域为$(1.645,+\infty)$

(4) 统计决策。

$Z=0.340\,093<Z_{0.05}=1.645$，所以不拒绝原假设，即男经理比女经理认为自己成功的人数比例高于 15%。

6.3.3 两个总体方差之比的假设检验

在对两个总体的方差进行比较时，通常将原假设与备择假设的基本形式表示成两个总体方差比值与数值 1 之间的比较关系。

构造两个总体的方差之比σ_1^2/σ_2^2的检验，其检验统计量要以其样本方差之比s_1^2/s_2^2为基础。其构造的过程如下。

两个总体是服从正态分布，其样本方差是服从χ^2分布，即有

$$\frac{(n_1-1)s_1^2}{\sigma_1^2}\sim\chi^2(n_1-1)\qquad\qquad\frac{(n_2-1)s_2^2}{\sigma_2^2}\sim\chi^2(n_2-1)$$

既然有两个χ^2分布，那么可以构造出一个F分布，有

$$F=\frac{\dfrac{(n_2-1)s_2^2}{\sigma_2^2}/(n_2-1)}{\dfrac{(n_1-1)s_1^2}{\sigma_1^2}/(n_1-1)}\sim F(n_2-1,n_1-1)$$

整理得

$$F=\frac{s_2^2}{s_1^2}\frac{\sigma_1^2}{\sigma_2^2}\sim F(n_2-1,n_1-1)$$

设两个总体方差之比的假设值为$\left(\dfrac{\sigma_1^2}{\sigma_2^2}\right)_0$，所以有其检验的统计量为

$$F=\frac{s_2^2}{s_1^2}\left(\frac{\sigma_1^2}{\sigma_2^2}\right)_0\sim F(n_2-1,n_1-1) \tag{6.19}$$

【例 6.16】 生产工序中的方差是工序质量的一个重要测度，通常较大的方差就意味着产品质量的波动程度大，需要通过寻找较小的工序方差来改进工序。现有一旧机器和一新机器两台机器，两台机器生产的袋装食品重量数据见表 6-15 所示。

表 6-15　两台机器生产的袋装食品重量数据

旧机器	2.95	3.45	3.50	3.75	3.48	3.26	3.33	3.20
	3.16	3.20	3.23	3.37	3.90	3.36	3.25	3.27
	3.20	3.22	2.98	3.45	3.70	3.34	3.18	3.35
	3.12	—	—	—	—	—	—	—
新机器	3.22	3.30	3.34	3.28	3.29	3.25	3.30	3.27
	3.38	3.34	3.35	3.19	3.35	3.05	3.36	3.28
	3.30	3.28	3.30	3.20	3.16	3.33	—	—

在显著性水平 $\alpha=0.05$ 下，检验新机器生产的袋装食品的重量与旧机器生产的袋装食品重要是否有显著的差异。

解：传统的假设检验步骤如下。

(1) 提出原假设和备择假设。

$$H_0:\frac{\sigma_1^2}{\sigma_2^2}=1；\ H_1:\frac{\sigma_1^2}{\sigma_2^2}\neq 1$$

(2) 构造检验的统计量，并计算其值。

$$F=\frac{s_2^2}{s_1^2}\left(\frac{\sigma_1^2}{\sigma_2^2}\right)_0\sim F(n_2-1,n_1-1)$$

其中$\left(\dfrac{\sigma_1^2}{\sigma_2^2}\right)_0=1$，根据样本的数据可得 $s_1{}^2=0.048\,808;\ s_2{}^2=0.005\,901$，则有

$$F=\frac{0.005\,901}{0.048\,808}=0.120\,9$$

(3) $\alpha=0.05$，确定拒绝原假设的区域为

$$(0,F_{1-\alpha/2}(21,24))\cup(F_{\alpha/2}(21,24),+\infty)$$

其中临界值查 F 分布表可得 $F_{0.975}(21,24)=0.422\,382$，$F_{0.025}(21,24)=2.310\,919$，所以拒绝域为 $(0,0.422\,382)\cup(2.310\,919,+\infty)$。

(4) 统计决策。

由于 $F=0.120\,9<F_{0.975}(21,24)=0.422\,382$，所以拒绝原假设，即新旧机器生产的袋装食品重量方差存在显著性差异。

6.4 案例分析：啤酒市场的调查与分析及 Excel 上机应用——啤酒印象与性别的相关性分析

在第 4 章的案例中，分析了性别对啤酒综合印象的影响，即分析男女两组的啤酒综合印象分数数据进行描述性分析，得出的结论是男性对啤酒的平均印象分数远高于女性，但当时还不能说性别对啤酒综合印象分数有影响。通过本章的学习，现在可以利用假设检验进行分析，分析性别是否对啤酒综合印象分数有显著性的影响。

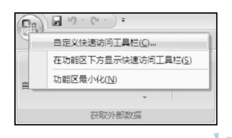


图 6.7 选择“自定义快速访问工具栏”选项

在分析性别是否对啤酒综合印象分数有显著性的影响前，需要安装数据分析工具库。安装的操作过程如下。

第一步：进入 Excel 界面，右击“Office 按钮”，在弹出的快捷菜单选择“自定义快速访问工具栏”选项，如图 6.7 所示，弹出“Excel 选项”对话框。

第二步：在“Excel 选项”对话框中，选择左侧的“加载项”选项，单击右下角“转到”按钮，如图 6.8 所示，弹出“加载宏”对话框。

第三步：在“加载宏”对话框中，选择“分析工具库”选项，单击“确定”按钮，等待安装数据分析工具库，如图 6.9 所示。



图 6.8 “Excel 选项”对话框

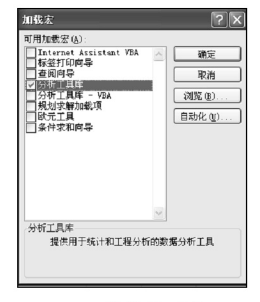


图 6.9 “加载宏”对话框

安装数据分析后，分析性别是否对啤酒综合印象分数有显著性的影响，即分析两个总体的均值是否相等，也就是分析两个总体均值之差是否等于 0。

第一步：提出原假设和备择假设

$$H_0:\mu_1-\mu_2=0；\quad H_1:\mu_1-\mu_2\neq 0$$

其中 μ_1 为女性总体的啤酒综合印象平均分数，μ_2 为男性总体对啤酒综合印象平均分数。

第二步：构造检验的统计量，并计算其值。

前面已经介绍过，检验两个总体均值之差是否等于 0，其检验的统计量要根据具体的情况，用不同的检验统计量。具体分为两大情况：一是两个样本相互独立；二是配对样本。经分析，该案例中的两个样本是相互独立的。

在两个样本相互独立的条件下，又分为两种情况：一是两个总体的方差已知；二是两个总体方差未知。那么该案例是属于两个总体方差未知情况，且是小样本，因为女性样本中有 11 人，男性样本中有 19 人。

在理论中，两个总体方差未知，且是小样本时，分为两种：一是两个总体方差相等；二是两个总体方差不等。所以在进行 t 检验前，要进行检验两个总体方差是否相等。

1. 检验两个总体方差是否相等

(1) 提出原假设和备择假设。

$$H_0:\sigma_1^2=\sigma_2^2；\quad H_1:\sigma_1^2\neq\sigma_2^2$$

(2) 构造检验的统计量，并计算其值。

根据前面的介绍，检验两个总体的方差是否相等，使用 F 检验，计算其值的操作过程如下。

第一步：打开“性别对啤酒印象分数的影响分析”工作表，单击“数据”→“分析”→“数据分析”按钮，弹出“数据分析”对话框，并在“分析工具”列表中选择“F-检验 双样本方差”选项，如图 6.10 所示。

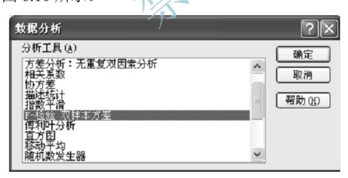


图 6.10 “数据分析”对话框

第二步：单击“确定”按钮，弹出“F-检验 双样本方差”对话框，在“变量 1 的区域”文本框中输入“$A$38:$A$50”，在“变量 2 的区域”文框中输入“$B$38:$B$56”，选择“标志”复选框，选中“输出区域”单选按钮，并在其文本框中输入“$E$51”，如图 6.11 所示。



图 6.11 “F-检验 双样本方差”对话框

第三步：单击“确定”按钮，得到如图 6.12 所示的统计结果。

F-检验 双样本方差分析		
	女	男
平均	3.166666667	9.166666667
方差	9.424242424	6.264705882
观测值	12	18
df	11	17
F	1.504339166	
P(F<=f) 单尾	0.217551324	
F 单尾临界	2.412561442	

图 6.12 统计结果

根据图 6.12，可知检验的统计量值为 F=1.504 339 166。

(3) 根据图 6.12，可知检验的统计量值 F 所对应的单尾 P 值为 0.217 551 324，而这里是双侧检验，应在单尾 P 值基础上乘以 2。

(4) 统计决策。

$P>0.05$，所以不拒绝原假设，即两个总体的方差是相等的。所以进行两个总体的均值之差检验时，要使用“t-检验：双样本等方差假设”。

构造两个总体均值之差的检验统计量使用“t-检验：双样本等方差假设”，即在“数据分析”对话框的“分析工具”列表中选择“t-检验：双样本等方差假设”选项后，单击“确定”按钮，如图 6.13 所示。



图 6.13 选择“t-检验：双样本等方差假设”选项

在“t-检验：双样本等方差假设”对话框的“变量 1 的区域”文本框中输入“$A$39：$A$50”，在“变量 2 的区域”文本框中输入“$A$39：$A$50”，“标志”复选框，选中“输出区域”单选按钮，并在其文本框中输入一个空白单元格，如输入“$E$63”，如图 6.14 所示。



图 6.14　“t-检验：双样本等方差假设”对话框

单击“确定”按钮，得到如图 6.15 所示的统计结果。

t-检验：双样本等方差假设		
	女	男
平均	3.166666667	9.166666667
方差	9.424242424	6.264705882
观测值	12	18
合并方差	7.505952381	
假设平均差	0	
df	28	
t Stat	-5.876443923	
P(T<=t) 单尾	1.27655E-06	
t 单尾临界	1.701130908	
P(T<=t) 双尾	2.5531E-06	
t 双尾临界	2.048407115	

图 6.15　统计结果

根据图 6.15，分析可得检验两个总体均值之差的检验统计量值为 $t=-4.8031$。

第三步：根据计算出的检验统计值，计算 P 值。

备择假设为两个总体均值之差不等于 0，所以此检验为双侧检验。根据图 6.15，可知双侧检验统计值所对应的 P 值 2.5531×10^{-6}。

第四步：统计决策。

$P=2.5531\times10^{-6}<0.01$，拒绝原假设，即两个总体均值之差不等 0，也就是说两个总体均值不相等，说明性别对啤酒综合印象分数有显著性的影响。

习　　题

一、单项选择题

1．某厂生产的零件直径服从正态分布，零件的直径标准要求为 10cm。某天质检人员从一批生产的零件中随机抽取 25 个，测得其直径的均值为 10.3cm，检验该批零件是否合格，则下列正确的假设形式是(　　)。

A．$H_0:\mu=10$；$H_1:\mu\neq10$　　B．$H_0:\mu\geqslant10$；$H_1:\mu<10$

C．$H_0:\mu\leqslant10$；$H_1:\mu>10$　　D．$H_0:\mu>10$；$H_1:\mu\leqslant10$

2．据有关部门估计该城市拥有汽车比例为 20%，然而有人认为这个比例实际上还要高，要检验该说法是否正确，则假设形式为(　　)。

A. $H_0:\pi=0.2$；$H_1:\pi\neq0.2$　　B. $H_0:\pi\leqslant0.2$；$H_1:\pi>0.2$

C. $H_0:\pi\geqslant0.2$；$H_1:\pi<0.2$　　D. $H_0:\pi<0.2$；$H_1:\pi\geqslant0.2$

3. 对总体参数的具体数值所做的陈述称为(　　)。

A. 假设　　B. 参数估计　　C. 假设检验　　D. 双侧检验

4. 利用样本的信息来检验总体参数的假设过程称为(　　)。

A. 假设　　B. 参数估计　　C. 假设检验　　D. 双侧检验

5. 备择假设通常是指研究人员(　　)。

A. 想收集证据要支持的观点　　B. 想收集证据要反对的观点

C. 想要支持的一个正确的观点　　D. 想要反对的一个正确的观点

6. 原假设通常是指研究人员(　　)。

A. 想收集证据要支持的观点　　B. 想收集证据要反对的观点

C. 想要支持的一个正确的观点　　D. 想要反对的一个正确的观点

7. 下列说法错误的是(　　)。

A. 在假设检验中，“=”总出现在原假设中

B. 在假设检验中，原假设和备择假设是完备事件组，且相互对立

C. 在假设检验中，原假设和备择假设只有一个成立

D. 在假设检验中，“=”既可以出现在原假设中，也可以出现在备择假设中

8. 在假设检验中，如果备择假设中出现“>”，则称为(　　)。

A. 单侧检验　　B. 右侧检验　　C. 左侧检验　　D. 双侧检验

9. 在假设检验中，如果备择假设中出现“<”，则称为(　　)。

A. 单侧检验　　B. 右侧检验　　C. 左侧检验　　D. 双侧检验

10. 在假设检验中，如果备择假设中出现“≠”，则称为(　　)。

A. 单侧检验　　B. 右侧检验　　C. 左侧检验　　D. 双侧检验

11.下列检验，属于双侧检验的是(　　)。

A. $H_0:\pi=0.2$；$H_1:\pi\neq0.2$　　B. $H_0:\pi\leqslant0.2$；$H_1:\pi>0.2$

C. $H_0:\pi\geqslant0.2$；$H_1:\pi<0.2$　　D. $H_0:\pi<0.2$；$H_1:\pi\geqslant0.2$

12. 下列检验，属于左侧检验的是(　　)。

A. $H_0:\pi=0.2$；$H_1:\pi\neq0.2$　　B. $H_0:\pi\leqslant0.2$；$H_1:\pi>0.2$

C. $H_0:\pi\geqslant0.2$；$H_1:\pi<0.2$　　D. $H_0:\pi<0.2$；$H_1:\pi\geqslant0.2$

13. 下列检验，属于右侧检验的是(　　)。

A. $H_0:\pi=0.2$；$H_1:\pi\neq0.2$　　B. $H_0:\pi\leqslant0.2$；$H_1:\pi>0.2$

C. $H_0:\pi\geqslant0.2$；$H_1:\pi<0.2$　　D. $H_0:\pi<0.2$；$H_1:\pi\geqslant0.2$

14. 在假设检验中，第Ⅰ类错误是指(　　)。

A. 实际上原假设是正确的，而拒绝了原假设

B. 实际上原假设是正确的，不拒绝原假设

C. 实际上原假设是错误的，拒绝原假设

D. 实际上原假设是错误的，而不拒绝原假设

15. 在假设检验中，第Ⅱ类错误是指(　　)。

A. 实际上原假设是正确的，而拒绝了原假设

B. 实际上原假设是正确的，不拒绝原假设

C. 实际上原假设是错误的，拒绝原假设

D. 实际上原假设是错误的，而不拒绝原假设

16．在假设检验中，样本容量不变的条件下，第Ⅰ类错误和第Ⅱ类错误的发生概率(　　)。

A．可以同时减小　　B．不能同时减小
C．可以同时增大　　D．只能同时增大

17．拒绝原假设的检验统计量的所有可能取值的集合，称为(　　)。

A．拒绝域　　B．双侧检验　　C．不拒绝域　　D．显著性水平

18．在假设检验中，对于犯第Ⅰ类错误的概率，我们称为(　　)。

A．显著性水平　　B．拒绝域　　C．置信水平　　D．不拒绝域

19．下列关于 P 值说法正确的是(　　)。

A．P 值越大，不拒绝原假设的可能性越大
B．P 值越小，拒绝原假设的可能性就越大
C．P 值越小，不拒绝原假设的可能性就越小
D．P 值越大，不拒绝原假设的可能性越小

20．在总体服从正态分布，方差已知的情况下，检验总体均值的统计量是(　　)。

A．$Z=\dfrac{\bar{x}-\mu_0}{\sigma/\sqrt{n}}$　　B．$Z=\dfrac{\bar{x}-\mu_0}{s/\sqrt{n}}$　　C．$t=\dfrac{\bar{x}-\mu_0}{s/\sqrt{n}}$　　D．$t=\dfrac{\bar{x}-\mu_0}{s^2/\sqrt{n}}$

21．在总体方差未知，大样本的情况下，检验总体均值的统计量是(　　)。

A．$Z=\dfrac{\bar{x}-\mu_0}{\sigma/\sqrt{n}}$　　B．$Z=\dfrac{\bar{x}-\mu_0}{s/\sqrt{n}}$　　C．$t=\dfrac{\bar{x}-\mu_0}{s/\sqrt{n}}$　　D．$t=\dfrac{\bar{x}-\mu_0}{s^2/\sqrt{n}}$

22．在总体服从正态分布，方差未知，小样本的情况下，检验总体均值的统计量是(　　)。

A．$Z=\dfrac{\bar{x}-\mu_0}{\sigma/\sqrt{n}}$　　B．$Z=\dfrac{\bar{x}-\mu_0}{s/\sqrt{n}}$　　C．$t=\dfrac{\bar{x}-\mu_0}{s/\sqrt{n}}$　　D．$t=\dfrac{\bar{x}-\mu_0}{s^2/\sqrt{n}}$

23．大样本的情况下，检验总体比例的统计量是(　　)分布。

A．标准正态　　B．t　　C．F　　D．χ^2

24．检验一个正态总体方差使用的统计量是(　　)分布。

A．标准正态　　B．t　　C．F　　D．χ^2

25．利用 P 值进行统计决策时，拒绝原假设的规则是(　　)。

A．$P<\alpha$　　B．$P>\alpha$　　C．$P=\alpha$　　D．$P=\alpha=0$

26．利用 P 值进行统计决策时，不拒绝原假设的规则是(　　)。

A．$P<\alpha$　　B．$P>\alpha$　　C．$P=\alpha$　　D．$P=\alpha=0$

二．简答题

1．简述传统假设检验的步骤。

2．在假设检验中，当样本容量一定的情况下，是否可以同时减小第Ⅰ类错误和第Ⅱ类错误的发生概率，为什么？

3．总结出不同情况的总体均值的假设检验统计量。

4．以大样本，$H_1:\pi\neq 40\%$ 为例，试写出其假设检验的过程。($\alpha=0.05$)

三．判断分析题

1．一个制造商想要检验新方法生产的零件直径是否比旧方法生产的零件直径方差(0.001 56mm^2)降低了，从新生产方法中随机抽取 100 个零件作为样本，测得零件的直径方差为 0.002 11mm^2。其检验的过程如下。($\alpha=0.05$)

(1) 提出原假设和备择假设。

$$H_0:\sigma^2\leqslant 0.001\,56；H_1:\sigma^2>0.001\,56$$

(2) 构造检验的统计量，并计算其值。

$$\chi^2=\frac{(n-1)S^2}{\sigma_0^2}=\frac{99\times 0.002\,11}{0.001\,56}=133.9$$

(3) $\alpha=0.05$，确定拒绝原假设的区域。

从备择假设中可以看出此处是右侧检验，所以拒绝域为$(\chi_\alpha{}^2(n-1),+\infty)$，其中$\chi_{0.05}{}^2(99)$查表得$\chi_{0.05}{}^2(99)=123.225\,2$，所以拒绝域为$(123.225\,2,+\infty)$。

(4) 统计决策。

$\chi^2=133.9>\chi_{0.05}{}^2(99)=123.225\,2$，所以拒绝原假设。

试判断这个制造商的检验过程是否正确。如果不正确，请写出正确的过程。

2．某研究人员为了检验总体均值是否大于的一个假设值 10，从总体方差未知中抽取一个样本容量为 25 的样本，测得其样本均值为 11，标准差为 5。其检验的过程如下。($\alpha=0.05$)

(1) 提出原假设和备择假设。

$$H_0:\mu\leqslant 10;\ H_1:\mu>10$$

(2) 构造检验的统计量，并计算其值。

$$Z=\frac{\overline{x}-\mu}{s/\sqrt{n}}=\frac{11-10}{5/\sqrt{25}}=1$$

(3) $\alpha=0.05$，确定拒绝原假设的区域。

备择假设中可以看出此处是右侧检验，所以拒绝域为$(Z_\alpha,+\infty)$，其中Z_α查表得$Z_{0.05}=1.645$，所以拒绝域为$(1.645,+\infty)$

(4) 统计决策。

$Z=1<Z_{0.05}=1.645$，所以不拒绝原假设。

试判断这个研究人员的检验过程是否正确。如果不正确，请出正确的过程。

四、计算题

1．一家汽车生产企业在广告中宣称“该公司的汽车可以保证在 2 年或 24 000km 内无事故”，但汽车的一个经销商认为保证“2 年”这一项是不必要的，因为该企业生产的汽车在 2 年内行驶的平均里程超过 24 000km。为了验证这一说法，该经销商随机抽取了 36 位顾客，测得这 35 位顾客在 2 年内行驶的里程均值为 245 00km，标准差为 600km，取显著性水平$\alpha=0.05$，对该问题进行假设检验。

2．一项研究发现，2013 年新购买小汽车的人中有 50%是女性，在 2014 年所做的一项调查中，随机抽取了 120 个新车主中有 71 人为女性，在$\alpha=0.01$的显著性水平下，检验 2014 年新车主中女性的比例是否显著性增加。

3．一项调查表明，2 年前每个大学生的上网平均时间为 6.7 h，而最近对 200 个大学生上网的时间调查结果是，每个大学生每天上网的平均时间为 7.25 h，标准差为 2.5 h，在显著性水平$\alpha=0.01$下，检验现如今大学生每天平均上网时间有无显著性的增加。

4．一项新型减肥方法宣称参加者在一个月内平均能减去 8g。为了验证该说法，某研究人员随机抽取了由 25 人使用该方法减肥的人组成样本，测其平均减重为 7kg，标准差为 3.2 kg。在$\alpha=0.01$的显著性水平下，检验这项减肥方法宣称的是否属实。

五、Excel 操作题

某企业为比较两种方法对员工进行培训的效果，采用方法 1 对 15 名员工进行培训，采用方法 2 对 12 名员工进行培训。培训后的测试分数见表 6-16 所示。

表 6-16　培训后的测试分数

方法 1	方法 2
56	59
51	57
42	53
47	52
50	57
43	68
41	54
56	59
52	53
50	54
46	62
48	57
47	—
45	—
46	—

在 $\alpha = 0.01$ 的显著性水平下，检验两种方法的培训效果是否存在显著性的差异。(提示：首先要检验两个总体的方差是否相等，之后再检验两种方法的培训效果是否存在显著性的差异)

第 7 章　方差分析

教学目标

1. 掌握方差分析的基本问题。
2. 掌握单因素方差分析。
3. 掌握方差分析中的多重比较。

导入案例

性别与某门课程成绩高低有关系吗?

某高校经济学专业的 3 名学生，决定研究性别与某门课程成绩高低是否有关系？如果存在关系的话，这种影响程度又如何呢？

调查的对象是该校凡学过此课程的几个专业的学生，样本的抽取方式是分层抽样与简单随机抽样结合，先根据年级划分层次，然后对各个班级作简单随机抽样，共抽取 150 名学生组成一个样本，然后对每个学生采用问卷调查。调查得到的男女学生成绩情况的汇总表见表 7-1 所示。

表 7-1　问卷调查汇总表

男	42	87	51	79	…	82
女	91	56	…	89		

其中，男生的样本量为 90 人，女生的样本量为 60 人。

这里涉及两个变量：一个是分类变量即性别；另一个是数值变量即成绩。根据表 7-1 的数据，你认为性别与成绩高低是否有关系呢？如何来检验两个变量之间是否存在关系呢？学完本章内容就很容易解决这样的问题。

这里以一个例题引出方差分析的定义。

【例 7.1】 设有 4 个总体，每个总体的均值分别为 μ_1，μ_2，μ_3，μ_4，试检验 4 个总体的均值是否相等。(显著性水平 $\alpha = 0.05$)

解：刚学完第 6 章的内容，所以用第 6 章的假设检验，但一般的假设检验一次只能检验两个总体均值是否相等，即需要检验 6 次，分别为：$\mu_1 = \mu_2$；$\mu_1 = \mu_3$；$\mu_1 = \mu_4$；$\mu_2 = \mu_3$；$\mu_2 = \mu_4$；$\mu_3 = \mu_4$。

很显然，这样的检验十分烦琐，同时还存在一个很大的问题，即 $\alpha = 0.05$，也就是说每次检验允许犯第 I 类错误的概率只是 0.05，如果检验 6 次，最后得出结论使我们犯第 I 类错误的概率将达到 $1-(1-\alpha)^6 = 0.265$，严重超过了题目中允许犯的第 I 类错误的概率，这是不允许的。这时就要使用方差分析。

7.1　方差分析的基本理论

7.1.1　方差分析的定义

根据例 7.1，可得方差分析定义。

定义 7.1　检验多个总体均值是否相等的统计方法，称为方差分析。

方差分析研究的目的是什么呢？为了回答这个问题，来看下面的例子。

【例 7.2】 某大学生毕业后，决定自主创业，开一家快餐店。在设计快餐店方案中，他想知道店的地理位置会不会影响他的营业额，若影响，影响的程度有多大，判断因地理位置产生的收益会不会大于成本；如果没有影响，在设计方案中就不用考虑地理位置了。为了得到结论，他进行了市场调查，分别随机的从不同地理位置抽取样本，所得数据见表 7-2 所示。

表 7-2　不同地理位置的超市营业额

单位：元

商业区	写字楼	居民区
41 000	18 000	26 500
30 500	29 000	31 000
45 000	33 000	22 000
38 000	22 000	29 000
31 000	17 000	35 000
39 000	25 600	30 000
59 000	29 000	44 500
48 000	28 300	48 000
51 000	26 000	—
47 000	24 600	—
41 500	—	—
39 000	—	—

试问，地理位置是否对营业收入有影响？

解：这名大学生把地理位置分为 3 类，分别为商业区、写字楼和居民区，所以说地理位置是分类数据，营业收入是数值数据，想知道分类数据是否对数值数据产生影响，即分类数据是自变量，数值数据是因变量。

如果商业区的快餐店平均营业收入 μ_1、写字楼区的快餐店平均营业收入 μ_2 和居民区的快餐店平均营业收入 μ_3 相等，那说明地理位置对营业收入没有影响；如果均值不全相等，则意味着地理位置对营业收入是有影响的。

通过上面的分析，要想研究地理位置对营业收入是否有影响，只需看 3 个总体的均值是否相等。而上面的数据是样本的数据，所以要用 3 组样本数据来推导 3 个总体均值是否相等，其中多个总体均值是否相等，称为方差分析。

也就是说，方差分析研究了分类型自变量对数值型因变量的影响，即方差分析采用的方法就是通过检验各总体的均值是否相等来判断分析类型自变量对数值型因变量是否有显著影响。

7.1.2　方差分析中的几个基本概念

定义 7.2　在方差分析中，所要检验的对象称为因素或因子(factor)。

定义 7.3　因素的不同表现称为水平或处理(treatment)。

定义 7.4　每个因子水平下得到的样本数据称为观测值。

例如，在例 7.2 中要分析地理位置对营业收入是否有显著影响。这里的“地理位置”是要检验的对象，称为“因素”或“因子”；商业区、写字楼和居民区是“地理位置”这一因素的具体表现，称为“水平”或“处理”；在每个地区下得到的样本数据(营业收入)，称为观测值。由于只涉及“地理位置”一个因素，因此称为单因素三水平的试验。

在只有一个因素的方差分析(称为单因素方差分析)中，涉及两个变量：①分类型自变量；②数值型的因变量。

定义 7.5　当方差分析中只涉及一个分类自变量时，称为单因素方差分析。

除了单因素方差分析之外，还有双因素的方差因析。

定义 7.6　当方差分析中涉及两个分类自变量时，称为双因素方差分析。

例如，在例 7.2 中，除了研究地理位置对营业收入影响外，还可能受其他分类数据的影响，如竞争对手的数量，把竞争对手的数量分为 0 个、1 个、2 个和 3 个及以上 4 个水平。这时就涉及两个分类自变量，称为双因素方差分析。

本书重点介绍单因素方差分析，双因素方差分析的原理与单因素方差分析的过程相似，不做介绍。

7.1.3　方差分析的基本思路

为了分析分类型自变量对数值型因变量的影响，需要从分析数据误差的来源入手。

先计算出 3 个总体下的样本均值，即有 $\bar{x}_1 = 42\,500$，$\bar{x}_2 = 25\,250$，$\bar{x}_3 = 33\,250$，从 3 个样本均值来看，商业区的样本均值高于居民区，居民区的营业收入又高于写字楼。但仅仅从样本均值上观察，还不能提供充分的证据证明不同地理位置对营业收入存在显著差异，因为这种差异可能是由抽样的随机性造成的。因此，需要有更准确的方法来检验这种差异是否显著，也就是要进行方差分析。

下面介绍方差分析的思路。

首先，注意到，例 7.2 中的所有观测值不同，存在差异，这种差异称为总误差。

其次，注意到在同一地理位置(同一个总体)下，样本的各观测值是不同的。例如，在商业区中，所抽取的 7 家快餐店的营业收入是不同的，这些数据之间存在差异，这种差异是组内误差。组内误差产生的原因：由于企业是随机抽取，因此它们之间的差异可以看成随机因素的影响造成的，或者说是由抽样的随机性所造成的，称为随机误差。

最后，在不同地理位置(不同总体)之间，各观测值也是不同的，即数据存在差异，称这种差异为组间误差。组间误差产生的原因：这种差异也可能是由抽样的随机性造成的，除此之外，还可能是由地理位置本身造成的，后者所形成的误差是由系统性因素造成的，称为系统误差。

从上面的分析中可以看出，组间误差和组内误差共同构成了总误差，如图 7.1 所示。

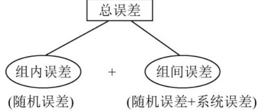


图 7.1　总误差、组间误差和组内误差的关系

如果不同地理位置对营业收入没有影响，那么组间误差只包含随机误差，而没有系统误差。如果不同地理位置对营业收入有影响，那么组间误差中就包含系统误差。即研究分类自变量(地理位置)对数值因变量(营业收入)是否有影响，转变为研究系统误差是否存在，即系统误差大小是否为 0。

想得到系统误差的大小值，只需计算出组内误差的大小和组间误差的大小，如果它们的比值接近 1，有随机误差≈随机误差+系统误差，即系统误差的大小为 0，不存在系统误差；反之，组间误差大小大于组内误差大小，它们之间的比值就会大于 1。当这个比值大到某种程度时，即存在系统误差，也就是说因素的不同水平之间存在显著差异，即自变量对因变量有影响。

定义 7.7　反映不同水平之间的数据误差的大小，称为组间平方和，记为 SSA。它反映不同水平之间的离散状况。

定义 7.8　反映全部数据的误差大小，称为总平方和，记为 SST。它反映全部数据的总离散状况。

定义 7.9　反映组内误差大小的平方和，称为组内平方和，记为 SSE。它反映了每个样本内各观测值的总离散状况。

7.1.4　方差分析的条件

方差分析中有 3 个基本的假定。

1. 每个总体都应服从正态分布

也就是说，对于因素的每一个水平，其观测值是来自正态分布总体的简单随机样本。

例如，在例 7.2 中，每个地理位置的营业收入必须服从正态分布。

2. 方差齐性

方差齐性是指每个总体的方差相同，也就是说，对于各组观察数据，它们是从具有相同方差的正态总体中抽取的。

例如，在例 7.2 中，每个地理位置的营业收入总体方差都相同。

3. 观测值是独立的

例如，在例 7.2 中，每个被抽中的快餐店的营业收入与其他的快餐店营业收入相互独立，没有任何关系。

设例 7.2 中的商业区的所有快餐店为总体为 X_1，则 $X_1 \sim N(\mu_1, \sigma^2)$；写字楼的所有快餐

店为总体 X_2，则 $X_2 \sim N(\mu_2, \sigma^2)$；居民区的所有快餐店为总体 X_3，则有 $X_3 \sim N(\mu_3, \sigma^2)$，且 3 组样本的观测值是相互独立的。

7.2 单因素方差分析

7.2.1 数据结构

在进行单因素方差分析时，需要得到下面的数据结构，为叙述方便，在单因素方差分析中，用 A 表示因素，因素的 k 个水平(总体)分别用 $A_1, A_2, \cdots, A_k$ 表示，每个观测值用 $x_{ij} (i=1,2,\cdots,k; j=1,2,\cdots,n)$ 表示，即 x_{ij} 表示第 i 个水平(总体)的第 j 个观测值。例如，x_{21} 表示第 2 个水平的第 1 个观测值。其中，从不同水平中抽取的样本容量可以相等，也可以不相等。每一水平下的样本容量为 n_i。

在例 7.2 中，地理位置是因素，因素有 3 个不同水平，即 $k=3$，商业区为 A_1，写字楼为 A_2，居民区为 A_3，其中商业区的样本容量为 $n_1=12$，写字楼的样本容量为 $n_2=10$，居民区的样本容量为 $n_3=8$。数据结构见表 7-3 所示。

表 7-3 单因素方差分析数据结构

商业区 A_1	写字楼 A_2	居民区 A_3
41 000(x_{11})	18 000(x_{21})	26 500(x_{31})
30 500(x_{12})	29 000(x_{22})	31 000(x_{32})
45 000(x_{13})	33 000(x_{23})	22 000(x_{33})
38 000(x_{14})	22 000(x_{24})	29 000(x_{34})
31 000(x_{15})	17 000(x_{25})	35 000(x_{35})
39 000(x_{16})	25 600(x_{26})	30 000(x_{36})
59 000(x_{17})	29 000(x_{27})	44 500(x_{37})
48 000(x_{18})	28 300(x_{28})	48 000(x_{38})
51 000(x_{19})	26 000(x_{29})	—
47 000(x_{110})	24 600(x_{210})	—
41 500(x_{111})	—	—
39 000(x_{112})	—	—

7.2.2 单因素方差分析的基本步骤

方差分析是检验自变量对因变量是否有显著影响，既然是检验就要满足假设检验的 4 个步骤。

(1) 提出原假设 H_0 和备择假设 H_1。

(2) 构造检验统计量，并计算其值。

(3) 根据给出的显著性水平 α，确定拒绝原假设的区域。

(4) 统计决策。

1. 提出原假设 H_0 和备择假设 H_1

确定方差分析的原假设 H_0 和备择假设 H_1 时，同样要从备择假设入手，备择假设是指研究人员予以支持的观点。

在前面已介绍，方差分析是检验多个总体均值是否相等的，研究人员认为分类自变量对数值因变量是有影响的，所以方差分析的原假设和备择假设内容如下。$H_0:\mu_1=\mu_2=\cdots=\mu_K$；$H_1:\mu_1,\mu_2,...,\mu_K$ 不全相等。原假设表明分类自变量对数值因变量没有显著影响；备择假设表明分类自变量对数值因变量有显著影响。其中，μ_i 为第 i 个总体的均值。

如果拒绝原假设，则意味着自变量对因变量有显著影响，也就是自变量与因变量之间有显著关系；如果不拒绝原假设，则没有证据显示自变量对因变量有显著影响，也就是不能认为自变量与因变量之间有显著关系。

注意：在备择假设中，要理解好“不全”，而不是“全不”。“不全”是指至少有两个总体的均值不相等就可以，而“全不”是指所有的总体均值之间全都不相等。

2. 构造检验的统计量，并计算其值

在第 6 章已经介绍过，检验的统计量首先要服从标准正态分布、χ^2 分布、t 分布和 F 分布 4 个分布之一，其次检验的统计量中不可以含有未知数，如果含有未知数，无法计算出检验统计量的数值。

如何构造这一统计量？这时，就要利用方差分析的思路来构造这一检验的统计量。

方差分析的思路：把研究分类自变量对数值因变量是否有影响，转为研究系统误差的大小是否为零，而要想计算出系统误差的大小，必须计算组间误差大小(组间平方和)和组内误差大小(组内平方和)，再将两者进行比值。

1) 计算误差大小

(1) 组内平方和。

组内平方和是衡量同一水平下数据差异大小的总误差，即求出方差分析中每一水平的误差大小之和，要想计算出每一水平的数据差异，首先要计算出每一水平的样本平均值。

① 计算因素各样本的均值。

设 $\bar{x}_i$ 为每一水平下的样本均值，则有每一水平的样本均值：

$$\bar{x}_i=\frac{\sum_{j=1}^{n_i}x_{ij}}{n_i}\quad i=1,2,\cdots,k \tag{7.1}$$

式中，n_i 为第 i 个总体的样本容量；x_{ij} 为第 i 个总体的第 j 个观测值。

② 计算每一水平下的组内误差大小。

设每一水平的组内误差大小为 SSE_i $(i=1,2,\cdots,k)$

$$\mathrm{SSE}_i=\sum_{j=1}^{n_i}(x_{ij}-\bar{x}_i)^2 \tag{7.2}$$

③ 计算组内平方和。

$$\mathrm{SSE}=\sum_{i=1}^{k}\mathrm{SSE}_i=\sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij}-\overline{x}_i)^2 \tag{7.3}$$

(2) 组间平方和。

组间平方和是衡量不同水平之间的数据差异大小。

① 计算全部观测值的总平均值。

全部观测值的总平均值是全部观测值的总和除以观测值的总个数。总平均值表示为$\overline{x}$。

$$\overline{x}=\frac{\sum_{i=1}^{k}\sum_{j=1}^{n_i}x_{ij}}{n}=\frac{\sum_{i=1}^{k}n_i\overline{x}_i}{n} \tag{7.4}$$

式中，$n=n_1+n_2+\cdots+n_k$。

② 计算组间平方和。

$$\mathrm{SSA}=\sum_{i=1}^{k}n_i(\overline{x}_i-\overline{x})^2 \tag{7.5}$$

(3) 总平方和。

总平方和是衡量所有观测数据的差异大小的。所以总平和的公式为

$$\mathrm{SST}=\sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij}-\overline{x})^2 \tag{7.6}$$

上述 3 个平方和之间的关系为

$$\mathrm{SST}=\mathrm{SSE}+\mathrm{SSR} \tag{7.7}$$

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij}-\overline{x})^2=\sum_{i=1}^{k}n_i(\overline{x}_i-\overline{x})^2+\sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij}-\overline{x}_i)^2$$

2) 构造检验统计量

由方差分析的思路可知，检验的统计量要从组间平方和和组内平方和两者的比值入手。

方差分析前提条件有 3 个：一是正态性，每个总体都要服从正态分布；二是方差齐性，每个总体的方差相同；三是观测值相互独立。

根据总体服从正态分布，方差齐性，可以推出每个总体下的样本均值均服从正态分布，又因为有

$$\mathrm{SSE}=\sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij}-\overline{x}_i)^2$$

$$\mathrm{SSA}=\sum_{i=1}^{k}n_i(\overline{x}_i-\overline{x})^2$$

$$\mathrm{SST}=\sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij}-\overline{x})^2$$

可推出 3 个平方和都是服从χ^2分布的，即有

$$\text{SSE} = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij} - \overline{x}_i)^2 \sim \chi^2(n-k)$$

$$\text{SSA} = \sum_{i=1}^{k} n_i(\overline{x}_i - \overline{x})^2 \sim \chi^2(k-1) \tag{7.8}$$

$$\text{SST} = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij} - \overline{x})^2 \sim \chi^2(n-1)$$

式中，k 为因素的不同水平的数目；n 为全部观测值的个数。

检验的统计量要从组间平方和和组内平方和的比值入手，而它们都是服从 χ^2 分布的，所以得到的检验统计量为

$$F = \frac{\text{SSA}/(k-1)}{\text{SSE}/(n-k)} \sim F(k-1, n-k) \tag{7.9}$$

定义 7.10　平方和除以相应的自由度，称为均方，用 MS 表示。

所以式(7.9)中的 $\dfrac{\text{SSA}}{k-1}$ 称为组间均方；$\dfrac{\text{SSE}}{n-k}$ 称为组内均方；即有

$$F = \frac{\text{SSA}/(k-1)}{\text{SSE}/(n-k)} = \frac{\text{MSA}}{\text{MSE}} \sim F(k-1, n-k) \tag{7.10}$$

3. 根据给出的显著性水平 α，确定拒绝原假设的区域

因为是右侧的单侧检验，所以拒绝原假设的区域为

$$(F_\alpha(k-1, n-k), +\infty)$$

4. 统计决策

当 $F > F_\alpha(k-1, n-k)$，检验的统计量落在拒绝原假设的区域中，所以拒绝原假设，即接受备择假设，意味着分类自变量对数值因变量是有影响的；当 $F \leqslant F_\alpha(k-1, n-k)$，没有落在拒绝原假设的区域中，不拒绝原假设，没有充足的证据证明分类自变量对数值因变量有影响。

【例 7.3】 沿用例 7.2，检验地理位置对快餐店的营业收入是否有影响。显著性水平 $\alpha = 0.05$。

解：

(1) 提出原假设 H_0 和备择假设 H_1。

$$H_0: \mu_1 = \mu_2 = \mu_3$$

$$H_1: \mu_1, \mu_2, \mu_3 \text{ 不全相等}$$

(2) 构造检验的统计量，并计算其值。

① 计算组内平方和。

$$\overline{x}_i = \frac{\sum_{j=1}^{n_i} x_{ij}}{n_i} \quad i = 1, 2, \cdots, k$$

$$\overline{x}_1 = 42\,500\text{；}\quad \overline{x}_2 = 25\,250\text{；}\quad \overline{x}_3 = 33\,250$$

$$\text{SSE} = \sum_{i=1}^{k}\text{SSE}_i = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij} - \overline{x}_i)^2$$

$$SSE = 7\,255\,000 + 22\,978\,500 + 5\,530\,000 = 15\,082\,850$$

② 计算组间平方和。

$$\bar{x} = \frac{\sum_{i=1}^{k}\sum_{j=1}^{n_i} x_{ij}}{n} = \frac{\sum_{i=1}^{k} n_i \bar{x}_i}{n} = 34\,283$$

其中，$n = n_1 + n_2 + n_2 = 30$

$$\begin{aligned} SSA &= \sum_{i=1}^{k} n_i (\bar{x}_i - \bar{x})^2 \\ &= 12 \times (42\,500 - 34\,283)^2 + 10 \times (25\,250 - 34\,283)^2 + 8 \times (33\,250 - 34\,283)^2 \\ &= 1\,634\,717\,000 \end{aligned}$$

③ 总平方和。

$$SST = \sum_{i=1}^{k}\sum_{j=1}^{n_i} (x_{ij} - \bar{x})^2$$

或

$$SST = SSE + SSR = 31\,430\,020$$

④ 检验的统计量为

$$\begin{aligned} F &= \frac{SSA/(k-1)}{SSE/(n-k)} \sim F(k-1, n-k) \\ &= \frac{1\,634\,717\,000/2}{1\,508\,285\,000/27} = \frac{81\,735.83}{5\,586.241} = 14.631\,63 \end{aligned}$$

(3) 根据给出的显著性水平 α，确定拒绝原假设的区域。

$$(F_\alpha(k-1, n-k), +\infty)$$

查 F 统计分布表得 $F_{0.05}(2,27) = 3.354\,131$，所以拒绝域为 $(3.354\,131, +\infty)$。

(4) 统计决策。

因为 $F = 14.631\,63 > F_{0.05}(2,27) = 3.354\,131$，落在拒绝原假设的区域中，所以拒绝原假设，即分类自变量对数值因变量有影响，也就是说，地理位置对快餐店的营业收入有影响。

在单因素方差分析中，也可以使用 P 值进行决策，这时只需改变方差分析中的第三步，根据得出的检验统计量值计算 P 值，把 P 值与显著性水平 α 相比，进行决策即可。

7.2.3 方差分析表

为了使方差分析的整个计算过程更加清晰，通常将上述过程列在一张表内，即方差分析表，其一般形式见表 7-4 所示。

表 7-4 方差分析表

误差来源	平方和 SS	自由度 df	均方 MS	F 值	P 值	F 临界值
组间	SSA	$k-1$	MSA	MSA/MSE	P 值	$F_\alpha(k-1, n-k)$
组内	SSE	$n-k$	MSE	—	—	—
总计	SST	$n-1$	—	—	—	—

则例 7.3 的计算结果可列成见表 7-5 所示的方差分析表。

表 7-5　不同地理位置的超市营业额方差分析表

误差来源	平方和 SS	自由度 df	均方 MS	F 值	P 值	F 临界值
组间	1 634 717 000	2	817 358 300	14.631 63	4.95878E-05	3.354 131
组内	1 508 285 000	27	55 862 410	—	—	—
总计	3 143 002 000	29	—	—	—	—

从上面的方差分析可以得出两种决策的方法：

(1) $F = 14.631\,63 > F_{0.05}(2,27) = 3.354\,131$，所以拒绝原假设。

(2) $P = 4.95878\text{E}-05 < \alpha = 0.05$，同样拒绝原假设。

7.2.4　关系强度的测量

例 7.3 的方差分析结果显示，不同地理位置的快餐店营业收入的均值有显著的差异，这意味着地理位置对快餐店的营业收入是有影响的，既然有影响，通常要知道地理位置对营业收入的影响程度。因为除了地理位置对快餐店营业收入有影响之外，还有其他因素影响营业收入，即想知道地理位置对营业收入影响占所有影响营业收入的因素的比例。

那么，如何量度它们之间的关系强度呢？可以用自变量平方和(SSA)及残差平方和(SSE)占总平方和(SST)的比例大小来反映。其中，自变量平方和占总平方和的比例记为 R^2，即

$$R^2 = \frac{\text{SSA}}{\text{SST}} \tag{7.11}$$

其平方根 R 就可以用来测量两个变量之间的关系强弱。如果 R 越大，代表关系强度越大；如果 R 越小，代表关系强度越小。

根据 $R^2 = \dfrac{\text{SSA}}{\text{SST}}$ 可知，其范围为 $[0,1]$。

【例 7.4】 沿用例 7.3 中的数据，计算地理位置对快餐店的影响程度有多大。

解：根据 $R^2 = \dfrac{\text{SSA}}{\text{SST}} = \dfrac{1\,634\,717\,000}{3\,143\,002\,000} = 0.52$

也就是说，地理位置对快餐店的营业收入的影响占总效应的 52%。

7.2.5　多重分析比较

通过上述的分析得出的结论是，不同地理位置的快餐店营业收入的均值是不相同的。但究竟哪些均值之间不相等呢？这种差异到底出现在哪些地理位置之间呢？也就是说，μ_1与μ_2，μ_1与μ_3，μ_2与μ_3之间究竟是哪两个均值不同呢？这就需要做进一步的分析，此时所用的方法就是多重比较方法。它是通过对总体均值之间的配对比较来进一步检验到底哪些均值之间存在差异。

多重比较方法有许多种，本书主要介绍由费希尔(Fisher)提出的最小显著差异方法，简记为 LSD。此方法主要检验哪两个均值不同，既然是检验，就要满足假设检验的 4 个步骤。

第一步：提出原假设和备择假设。

从备择假设入手，所以提出的原假设和备择假设内容如下：

$$H_0: \mu_i = \mu_j;\quad H_1: \mu_i \neq \mu_j \qquad (i \neq j;\ i,j = 1,2,3)$$

第二步：构造检验统计量，并计算其值。

检验的统计量从检验内容的样本入手，所以其检验的统计量如下：

$$\bar{x}_i - \bar{x}_j$$

第三步：根据给出的显著性水平 α 的数值，确定拒绝原假设的区域。

由第一步可得出，该检验为双侧检验，则费希尔的拒绝原假设区域为

$$(-\infty,-\text{LSD})\cup(\text{LSD},+\infty)$$

其中 LSD 的计算公式为

$$\text{LSD}=t_{\alpha/2}(n-k)\sqrt{\text{MSE}(\frac{1}{n_i}+\frac{1}{n_j})} \tag{7.12}$$

$t_{\alpha/2}(n-k)$ 为 t 分布的临界值，可以通过查 t 分布表得到，其自由度为 $n-k$，其中 k 为因素中水平的个数；MSE 为组内均方；n_i、n_j 是第 i 个样本和第 j 个样本的样本容量。

第四步：统计决策。

如果第二步的检验统计量落在第三步中的拒绝域内，则拒绝原假设；否则，不拒绝原假设。

【例 7.5】 沿用例 7.3 中的数据，试找出哪些地理位置的快餐店营业收入均值不相同。(显著性水平 $\alpha=0.05$)

解：

(1) 提出原假设和备择假设。

检验 1：$H_0:\mu_1=\mu_2$； $H_1:\mu_1\neq\mu_2$

检验 2：$H_0:\mu_1=\mu_3$； $H_1:\mu_1\neq\mu_3$

检验 3：$H_0:\mu_2=\mu_3$； $H_1:\mu_2\neq\mu_3$

(2) 构造检验的统计量，并计算其值。

检验 1：$\bar{x}_1-\bar{x}_2=42\,500-25\,250=17\,250$

检验 2：$\bar{x}_1-\bar{x}_3=42\,500-33\,250=9\,250$

检验 3：$\bar{x}_2-\bar{x}_3=25\,250-33\,250=-8\,000$

(3) 根据给出的显著性水平 α 的数值，确定拒绝原假设的区域

$$(-\infty,-\text{LSD})\cup(\text{LSD},+\infty)$$

检验 1：$\text{LSD}=t_{\alpha/2}(n-k)\sqrt{\text{MSE}\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}$

$$=t_{0.025}(27)\sqrt{55\,862\,410\left(\frac{1}{12}+\frac{1}{10}\right)}$$

$$=2.373\,417\,186\times\sqrt{55\,862\,410\times\left(\frac{1}{12}+\frac{1}{10}\right)}$$

$$=7\,595.47$$

所以拒绝域为 $(-\infty,-7\,595.47)\cup(7\,595.47,+\infty)$ 。

检验 2：$\text{LSD}=t_{\alpha/2}(n-k)\sqrt{\text{MSE}\left(\frac{1}{n_1}+\frac{1}{n_3}\right)}$

$$= 2.373\,417\,186 \times \sqrt{55\,862\,410 \times \left(\frac{1}{12} + \frac{1}{8}\right)}$$

$$= 8\,096.80$$

所以拒绝域为 $(-\infty, -8\,096.80) \cup (8\,096.80, +\infty)$。

检验 3： $\text{LSD} = t_{\alpha/2}(n-k)\sqrt{\text{MSE}\left(\frac{1}{n_2} + \frac{1}{n_3}\right)}$

$$= 2.373\,417\,186 \times \sqrt{55\,862\,410 \times \left(\frac{1}{10} + \frac{1}{8}\right)}$$

$$= 84.144\,4$$

所以拒绝域为 $(-\infty, -8\,414.44) \cup (8\,414.44, +\infty)$。

(4) 统计决策。

检验 1： $\bar{x}_1 - \bar{x}_2 = 17\,250 > 7\,595.47$，所以拒绝原假设，即 $\mu_1 \neq \mu_2$。

检验 2： $\bar{x}_1 - \bar{x}_3 = 9\,250 > 8\,096.80$，所以拒绝原假设，即 $\mu_1 \neq \mu_3$。

检验 3： $\bar{x}_2 - \bar{x}_3 = -80 > -8\,414.44$，所以不拒绝原假设，即 $\mu_2 = \mu_3$。

根据以上的分析，最后该大学生选择的地理位置如果是商业区，则他的营业收入会达到最大。

7.3　案例分析：啤酒市场的调查与分析及 Excel 上机应用——啤酒印象与学历的相关性分析

在第 4 章的案例中，分析了学历对啤酒综合印象的影响，即分析高中及以下、大专、本科和研究生及以上 4 组的啤酒综合印象分数数据进行描述性分析，得出的结论是，高中及以下啤酒的平均印象分数最高，其次是研究生及以上的平均分数，再次是大专的平均分数，最小的是本科的平均分数，但 4 组样本平均分数还不能说明学历对啤酒综合印象分数有影响。通过学习本章，现在可以利用方差分析进行分析，分析学历是否对啤酒综合印象分数有显著性的影响。分析的过程如下。

第一步：提出原假设和备择假设。

$$H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4$$

$$H_1: \mu_1, \mu_2, \mu_3, \mu_4 \text{ 不全相等}$$

其中，μ_1 为高中及以下总体的啤酒综合印象平均分数；μ_2 为大专总体的啤酒综合印象平均分数；μ_3 为本科总体的啤酒综合印象平均分数；μ_4 为研究生及以上总体的啤酒综合印象平均分数。

第二步：构造检验的统计量，并计算其值。

根据前面的介绍，已知检验的统计量为 F 检验，计算 F 的值软件操作过程如下。

(1) 打开“学历对啤酒印象分数的影响分析”工作表，单击“数据”→“分析”→“数据分析”按钮，弹出“数据分析”对话框，如图 7.2 所示。

(2) 在“数据分析”对话框中的“分析工具”列表中选择“方差分析：单因素方差分

析”选项，单击“确定”按钮，弹出“方差分析：单因素方差分析”对话框，在“输入区域”文本框中输入“$A$40：$D$56”，选择“标志位于第一行”复选框，选中“输出区域”单选按钮，并在其文本框中输入“$F$65”，如图 7.3 所示。



图 7.2 “数据分析”对话框

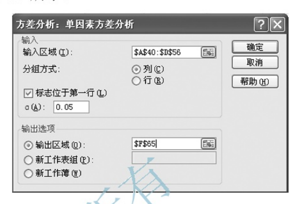


图 7.3 “方差分析：单因素方差分析”对话框

(3) 单击“确定”按钮，输出如图 7.4 所示的统计结果。

根据图 7.4 的分析结果，可得到检验的统计量 F 的值为$1.387\,22$。

第三步：确定拒绝原假设的区域，或计算 P 值。

根据图 7.4 的结果可知，拒绝原假设的区域为$(2.975\,154,+\infty)$，或$P=0.268\,845$。

方差分析：单因素方差分析						
SUMMARY						
组	观测数	求和	平均	方差		
高中及以下	2	21	10.5	4.5		
大专	6	44	7.3333333	19.466667		
本科	16	89	5.5625	15.4625		
研究生及以上	6	49	8.1666667	14.166667		
方差分析						
差异源	SS	df	MS	F	P-value	F crit
组间	64.7625	3	21.5875	1.38722	0.26884531	2.975154
组内	404.60417	26	15.561699			
总计	469.36667	29				

图 7.4 统计结果

第四步：统计决策。

以传统的假设检验方法决策，有$F=1.387\,22<2.975\,154$，所以不拒绝原假设，即学历对啤酒综合印象分数没有显著性的影响。

如果使用 P 值进行决策，$P=0.268\,845>0.1$，所以不拒绝原假设，得到相同的结论。

习　　题

一、单项选择题

1．与假设检验方法相比，方差分析方法可以使犯第 I 类错误的概率(　　)。

A．提高　　B．降低　　C．等于 0　　D．等于 1

2. 方差分析是检验(　　)。

A. 多个总体方差是否相等的统计方法

B. 多个总体均值是否相等的统计方法

C. 多个样本方差是否相等的统计方法

D. 多个样本均值是否相等的统计方法

3. 在方差分析中，所提出的原假设是 $H_0:\mu_1=\mu_2=\cdots=\mu_k$，备择假设是(　　)。

A. $H_1:\mu_1,\mu_2,\cdots,\mu_k$ 全不相等　　B. $H_1:\mu_1>\mu_2>\cdots>\mu_k$

C. $H_1:\mu_1<\mu_2<\cdots<\mu_k$　　D. $H_1:\mu_1,\mu_2,\cdots,\mu_k$ 不全相等

4. 在方差分析中，进行多重比较的前提是(　　)。

A. 拒绝原假设

B. 不拒绝原假设

C. 可以拒绝原假设也可以不拒绝原假设

D. 各样本均值相等

5. 在方差分析中，当结论为拒绝原假设时，则意味着(　　)。

A. 分类自变量对数值因变量有显著性的影响

B. 分类自变量对数值因变量没有显著性的影响

C. 多个总体均值中至少有一对均值不等

D. 多个总体均值之间全不相等

6. 方差分析中的检验统计量是(　　)分布。

A. 标准正态　　B. t　　C. F　　D. χ^2

7. 在方差分析中，检验统计量是(　　)。

A. $F=\dfrac{\text{SSA}/k-1}{\text{SSE}/n-k}$　　B. $F=\dfrac{\text{SSA}/k}{\text{SSE}/n-k}$

C. $F=\dfrac{\text{SSA}/k-1}{\text{SST}/n-1}$　　D. $F=\dfrac{\text{SSA}/k-1}{\text{SST}/n-k}$

8. 在方差分析中，拒绝原假设的区域为(　　)。

A. $(F_\alpha(k-1,n-k),+\infty)$　　B. $(F_{1-\alpha}(k-1,n-k),+\infty)$

C. $(0,F_{1-\alpha}(k-1,n-k))$　　D. $(-\infty,F_{1-\alpha}(k-1,n-k))$

9. 在方差分析中，组间平方和、组内平方和、总平方和的自由度分别为(　　)。

A. $k-1,n-k,n-1$　　B. $k-1,n-1,n-k$

C. $n-1,k-1,n-k$　　D. $n-1,n-k,k-1$

10. 在方差分析中，涉及一个分类的自变量，称为(　　)。

A. 单因素方差分析　　B. 双因素方差分析

C. 可重复双因素方差分析　　D. 不可重复的双因素方差分析

11. 最小显著差异方法是寻找哪些总体均值不等的方法，其原假设和备择假设为(　　)。

A. $H_0:\mu_i=\mu_j$；$H_1:\mu_i\neq\mu_j(i\neq j)$　　B. $H_0:\mu_i\leqslant\mu_j$；$H_1:\mu_i>\mu_j(i\neq j)$

C. $H_0:\mu_i\geqslant\mu_j$；$H_1:\mu_i<\mu_j(i\neq j)$　　D. $H_0:\mu_i>\mu_j$；$H_1:\mu_i\leqslant\mu_j(i\neq j)$

12. 最小显著差异方法的检验统计量是(　　)分布。

A. 标准正态　　B. t

C. F　　D. χ^2

13．最小显著差异值的计算公式是(　　)。

A．$\text{LSD}=t_{\alpha/2}(n-1)\sqrt{\text{MSE}\left(\frac{1}{n_i}+\frac{1}{n_j}\right)}$

B．$\text{LSD}=t_{\alpha/2}(n-k)\sqrt{\text{MSE}\left(\frac{1}{n_i}+\frac{1}{n_j}\right)}$

C．$\text{LSD}=t_{\alpha/2}(n-k)\sqrt{\text{MSA}\left(\frac{1}{n_i}+\frac{1}{n_j}\right)}$

D．$\text{LSD}=t_{\alpha/2}(k-1)\sqrt{\text{MSE}\left(\frac{1}{n_i}+\frac{1}{n_j}\right)}$

14．在方差分析中，总平方和(SST)、组间平方和(SSA)、组内平方和(SSE)的关系为(　　)。

A．SST=SSA+SSE　　B．SSA=SST+SSE

C．SSE=SST+SSA　　D．SST=SSA−SSE

15．在方差分析中，衡量分类自变量对数值因变量的影响程度指标是(　　)。

A．R^2　　B．SST　　C．SSA　　D．SSE

二、简答题

1．简述方差分析的思路。

2．简述方差分析的几个基本的假定。

3．简述方差分析的步骤。

4．简述最小差异方法的步骤。

三、计算题

1．某家企业采用自动生产线罐装饮料，要求每罐的容量为255mL，现有4种自动生产线，为了检验每种生产线是否有显著的差异，随机从各生产线抽取一组样本，数据见表7-6所示。

表7-6　生产线数据

单位：mL

生产线1	生产线2	生产线3	生产线4
256	261	254	249
260	263	253	248
245	258	251	251
241	249	252	256
251	261	257	257
253	251	248	258
—	249	249	249
—	248	—	248
—	256	—	—

取显著性水平$\alpha=0.01$，检验4个生产线的装填量是否有显著性差异。

2．城市道路交通管理部门为了研究不同的路段对行车是否有影响，让一名交通警察分别在3个路段亲自驾车进行实验，通过实验共获得15个行车时间的数据，见表7-7所示。

表 7-7　行车时间数据

单位：min

路段 1	路段 2	路段 3
36.5	28.1	32.4
34.1	29.9	33.0
37.2	32.2	36.2
35.6	31.5	35.5
—	30.1	35.1
—	38.0	—

取显著性水平 $\alpha = 0.01$，检验 3 个路段对行车时间否有显著性影响。

3．城市道路交通管理部门为了研究不同的时间段对行车是否有影响，让一名交通警察分别在 3 个时间段亲自驾车进行实验，通过实验共获得 30 个行车时间的数据，通过对每个时间段的行车时间进行方差分析得到表 7-8 所示的方差分析表。

表 7-8　方差分析表

差异源	SS	df	MS	*F*	*F*-crit
组间			210		3.354 131
组内	3 836			—	—
总计			—	—	—

(1) 完成上面的方差分析表。

(2) 若显著性水平 $\alpha = 0.05$，检验 3 个时间段的行车时间是否有显著性的差异。

四、Excel 操作题

1．利用 Excel 操作，检验第三题的第一题的 4 个生产线的装填量是否有显著性差异，并进行方差分析。

2．利用 Excel 操作，检验第三题的第二题的 3 个路段对行车时间否有显著性影响，并进行方差分析。

第 8 章 相关与一元回归分析

教学目标

1. 掌握变量相关关系分析。
2. 掌握一元回归分析，包括参数估计的方法(最小二乘法)、线性关系检验、回归系数的检验。
3. 掌握一元回归预测。
4. 掌握多元回归分析。
5. 掌握回归分析的软件操作。

本章主要介绍相关与回归分析，相关与回归是处理变量之间关系的一种统计方法。从所处理的变量多少来看，如果研究的是两个变量之间的关系，称为简单相关与简单回归分析；如果研究的是两个以上变量之间的关系，称为多元相关与多元回归分析。从变量之间的关系形态上看，有线性相关与线性回归分析及非线性相关与非线性回归分析。其中多元回归分析在本章 8.4 节介绍，重点介绍一元线性回归分析，因为多元线性回归分析原理同一元线性回归相同。

本章主要目的是利用相关与回归分析进行经济预测和经济控制。要达到此目的，需分 3 步进行。

第一步：进行相关分析。目的是判断因变量和自变量之间是否具有线性关系。

注意：这里把关系定为线性关系，因为非线性的关系，可以通过转换为线性关系。

第二步：回归分析。如果第一步判断出变量之间存在线性关系，则要进行变量的回归分析。

第三步：经济预测和经济控制。这一步主要是利用第二步的回归分析，进行经济预测和经济控制。

其中第一步是 8.1 节的内容；第二步是 8.2 节的内容；第三步是 8.3 节的内容。

8.1　相关分析的基本理论

8.1.1　变量间的关系

在生产和经营活动中，经常要对变量之间的关系进行分析。例如，在企业生产中，要对影响生产成本的各种因素进行分析，以达到控制成本的目的；在商业活动中，需要分析广告费支出与销售量之间的关系，进而通过广告费支出来预测销售量等。统计分析的目的在于，根据统计数据确定变量之间的关系形态及其关联程度，探索出其内在的数量规律性。

人们在实践中发现，变量之间的关系形态可分为两种类型，即函数关系和相关关系。其中函数关系是人们比较熟悉的。设有两个变量 x 和 y，变量 y 随变量 x 一起变化，并完全依赖于 x，当变量 x 取某个值时，y 依确定的关系取相应的值，则称 y 是 x 的函数，记为 $y=f(x)$。其中，x 称为自变量，y 称为因变量。函数关系是一一对应的确实关系，但在实际问题中，变量之间的关系往往不是那么简单。

例如，考察居民储蓄与居民家庭收入这两个变量，它们之间就不存在完全确定的关系。也就是说，收入水平相同的家庭，其储蓄额往往不同；反之，储蓄额相同的家庭，其收入水平也可能不相同。可见，家庭储蓄并不能完全由家庭收入的所确定，因为家庭收入尽管与家庭储蓄有密切的关系，但它并不是影响储蓄的唯一因素，还有银行利率、消费水平等其他因素的影响作用。正是由于影响一个变量的因素非常多，才造成了变量之间的关系的不确定性。

定义 8.1　变量之间存在的不确定的数量关系，称为相关关系。

8.1.2　相关分析

相关分析就是对两个变量之间线性关系的描述与度量，通常状况下，两个变量的总体数据是不易得到的，这时通常使用推断统计，利用两个变量的样本推导总体的关系，即相关分析要解决的问题包括：①两变量的样本之间是否存在线性的关系；②两变量的样本之间的关系强度如何；③样本所反映的变量之间的关系能否代表总体变量之间的关系。

为解决这些问题，在进行相关分析时，应对总体做一个基本假定，即两个变量都是随机变量。

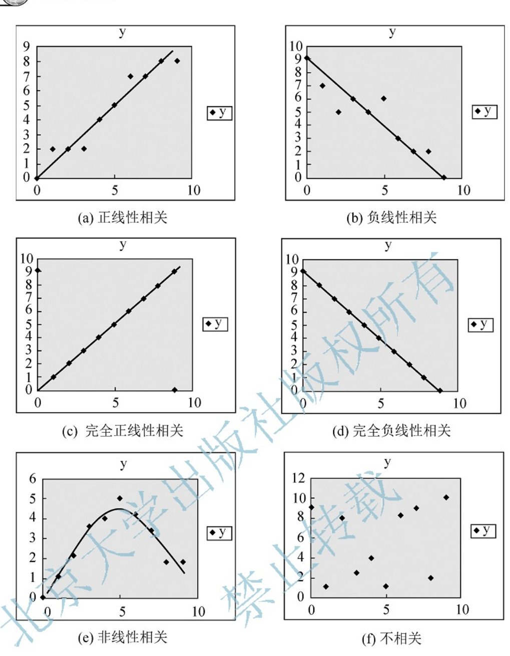
下面按照以上的相关分析的问题一一展开讲解。

1. 散点图

通常使用散点图来判断两变量的样本之间是否存在线性关系。

定义 8.2　对于两个变量 x 和 y，通过观察或试验可以得到若干组数据，记为 (x_i, y_i)， $i=1,2,\cdots,n$。如果用坐标的水平轴代表自变量 x，用纵轴代表因变量 y，那么每组数据 (x_i, y_i) 在坐标系中就可用一个点表示，n 组数据在坐标系中形成的 n 个点称为散点，由坐标及其散点形成的二维数据图称为散点图。

不同形态的散点图如图 8.1 所示。



(a) 正线性相关

(b) 负线性相关

(c) 完全正线性相关

(d) 完全负线性相关

(e) 非线性相关

(f) 不相关

图 8.1　不同形态的散点图

从图 8.1 中可以看出，(a)和(c)呈现出两个变量正相关，(b)和(d)呈现出两个变量是负相关，(e)是呈现出非线性相关，(f)不相关。

如果两个变量的样本数据表现出是上面的(a)、(b)、(c)和(d)任意一种，则两变量的样本之间存在线性关系。

【例 8.1】 为了研究所得产量与生产费用支出之间的关系，某汽车商管理部门随机抽取了 12 家汽车生产企业，得到它们的产量与生产费用支出的数据，见表 8-1 所示。绘制产量与生产费用的散点图，判断二者之间的关系形态。

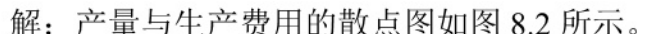


表 8-1　12 家汽车生产企业的产量与生产费用数据

企业编号	产量/台	生产费用/万元	企业编号	产量/台	生产费用/万元
1	40	130	7	84	165
2	42	150	8	100	170
3	50	155	9	116	167
4	55	140	10	125	180
5	65	150	11	130	175
6	78	154	12	140	185

解：产量与生产费用的散点图如图 8.2 所示。

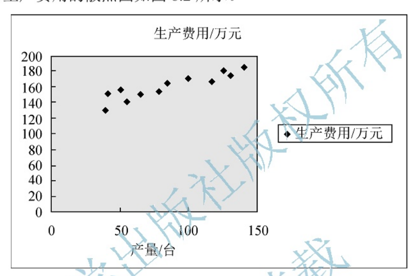


图 8.2　产量与生产费用的散点图

从图 8.2 中可以看出，随着产量不断地增加，生产费用也越大，二者的数据点分布在一条直线的附近，因此二者之间具有线性相关关系。

注意：此时的二者具有的线性相关关系是指二者的样本，并不是指总体。

2. 相关系数

通过散点图可以判断两个变量之间有无相关关系，只能对变量间的关系形态做出大致的描述，要想知道两个变量的关系强度，需要计算相关系数。

《概率与数理统计》已经介绍过，如果是计算总体的相关系数，记为 ρ；如果计算样本相关系数，记为 γ。在上面已经介绍过，总体的数据是不易得到的，通常只能得到样本的数据，所以这里所说的相关系数是指样本的。样本的相关系数又称皮尔逊相关系数，则样本相关系数的计算公式为

$$\gamma = \frac{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)}{\sqrt{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2\left(y_i - \bar{y}\right)^2}}$$

为了根据原始数据计算 γ，可以推导出的简化计算公式为

$$\gamma = \frac{\sum x_i y_i - n\overline{xy}}{\sqrt{\sum x_i^2 - n\bar{x}^2}\sqrt{\sum y_i^2 - n\bar{y}^2}}$$

样本相关系数γ具有以下特点。

(1) γ的取值范围为$-1 \leqslant \gamma \leqslant 1$，如果$0 < \gamma \leqslant 1$，则表明两个变量存在正线性相关关系；如果$-1 \leqslant \gamma < 0$，则表明两变量存在负线性相关关系；如果$\gamma = 0$时，说明两个变量之间不存在线性相关关系。

(2) γ具有对称性。γ的具有对称性是指x与y之间的相关系数γ_{xy}和y与x之间的相关系数γ_{yx}相等，即$\gamma_{xy} = \gamma_{yx}$。

(3) γ的大小与x和y的原点及尺度无关。例如，研究儿童身高与年龄的关系分析，无论身高采用 m 还是 cm 作单位，都不会改变身高和年龄的相关系数数值。

(4) γ仅仅是x与y之间线性关系的一个度量，它不能用于描述非线性关系。样本的相关系数这一点决定了，在进行相关分析，首先要判断两变量是否存在线性关系，如果存在线性关系，才能计算关系强度；否则，不可以计算相关系数。

(5) γ虽然是两个变量之间线性关系的一个度量，却不一定意味着x与y一定存在因果关系。

【例 8.2】 沿用例 8.1 中的数据，计算产量与生产费用的线性关系强度。

解：样本相关系数的公式为

$$
\begin{aligned}
\gamma &= \frac{\sum x_i y_i - n\bar{x}\bar{y}}{\sqrt{\sum x_i^2 - n\bar{x}^2}\sqrt{\sum y_i^2 - n\bar{y}^2}} \\
&= \frac{170\,094 - 164\,085.4}{\sqrt{101\,835 - 87\,552.08} \times \sqrt{310\,505 - 307\,520.1}} \\
&= \frac{6\,008.583}{\sqrt{14\,282.92} \times \sqrt{2\,984.917}} \\
&= \frac{6\,008.583}{119.511\,2 \times 54.634\,39} = \frac{6\,008.583}{6\,529.419} = 0.920\,232
\end{aligned}
$$

由上面的计算结果可得到产量与生产费用的关系强度为 0.920 232，具有较强的正线性相关关系。

可以利用 Excel 中的相关系数函数计算相关系数，操作过程如下。

进入 Excel 界面，单击“插入函数”按钮，弹出“插入函数”对话框。单击“或选择类别”的下拉按钮，在弹出的下拉列表中选择“统计”选项，并在“选择函数”列表中选择 CDRREL 选项，然后单击“确定”按钮，弹出“函数参数”对话框。在对话框中输入两组要计算的数据区域，单击“确定”按钮，返回结果相关系数为 0.920 232 426，如图 8.3 所示。



图 8.3 “函数参数”对话框

3. 相关关系的显著性检验

从上面的散点图和相关系数的计算，得到两变量的样本具有较强的正线性相关关系，如果想知道两变量的总体是否具有较强的正线性相关关系，采用了推断统计，来检验样本所反映的变量之间的关系能否代表总体变量之间的关系。如何去检验？这里使用的方法是相关关系的显著性检验。既然是检验，那么应满足假设检验的 4 个步骤。

(1) 提出原假设和备择假设。

研究人员希望是两个变量存在较强的线性相关关系，而且前面已介绍过，总体相关系数用 ρ 表示，即研究人员希望 $\rho \neq 0$。

$$H_0:\rho=0，H_1:\rho\neq 0$$

(2) 构造检验的统计量，并计算其值。

检验的统计量构造要从其样本入手，即从 γ 来构造。通常采用费尔希提出的 t 分布检验，该检验可以用于小样本，也可以用于大样本。

$$t=|\gamma|\sqrt{\frac{n-2}{1-\gamma^2}}\sim t(n-2) \tag{8.1}$$

(3) 根据给出的显著性水平 α，确定拒绝原假设的区域。

$$(-\infty,-t_{\alpha/2}(n-2))\cup(t_{\alpha/2}(n-2),+\infty)$$

通过查 t 分布表，可查出 $t_{\alpha/2}(n-2)$ 的临界值。

(4) 若 $|t|>t_{\alpha/2}(n-2)$，则拒绝原假设，表明总体的两个变量之间存在显著的线性关系。若 $|t|<t_{\alpha/2}(n-2)$，则不拒绝原假设，表明总体的两个变量之间不存在显著的线性关系。

【例 8.3】 沿用例 8.1 中的数据，检验产量与生产费用(总体)之间的相关关系是否显著。(显著性水平 $\alpha=0.05$)

解：

(1) 提出原假设和备择假设。

$$H_0:\rho=0，H_1:\rho\neq 0$$

(2) 构造检验的统计量，并计算其值。

$$\begin{aligned}t&=|\gamma|\sqrt{\frac{n-2}{1-\gamma^2}}\sim t(n-2)\\&=|0.920\,232|\sqrt{\frac{12-2}{1-0.920\,232^2}}\\&=0.920\,232\times\sqrt{65.285\,96}\\&=0.920\,232\times 8.079\,973=7.435\,453\end{aligned}$$

(3) 根据给出的显著性水平 $\alpha=0.05$，确定拒绝原假设的区域。

$$(-\infty,-t_{0.025}(10))\cup(t_{0.025}(10),+\infty)$$

通过查 t 分布表，可查出 $t_{0.025}(10)=2.633\,767$ 的临界值。

(4) 因为 $|t|=7.435\,453>t_{0.025}(10)=2.633\,767$，所以拒绝原假设，表明样本所反映的变量之间的关系能否代表总体变量之间的关系，即产量与生产费用两个变量存在线性关系。

8.2 一元线性回归分析

8.2.1 回归分析的含义

既然由相关分析得出两个变量之间存在线性关系，那么下面就要考察变量之间的数量伴随关系，并通过一定的数学表达式将这种关系描述出来，即回归分析。

这里要强调一下，要对两个变量的总体进行回归分析，而两个变量的总体数据是不易收集的，同样采用推断统计，利用样本数据推出总体的回归模型。具体来说，回归分析主要解决以下几个方面的问题。

(1) 从一组样本数据出发，确定出变量之间的数学关系式(样本的回归模型)。

(2) 对样本回归模型进行评价。

(3) 检验样本的回归模型是否能代表总体的回归模型，即对样本关系式的进行各种统计检验。

8.2.2 一元线性回归模型

1. 总体回归模型

定义 8.3　在回归分析中，被预测或被解释变量，称为因变量，用 y 表示。

定义 8.4　在回归分析中，用来预测或解释因变量的一个或多个变量，称为自变量，用 x 表示。

例如，例 8.1 中产量与生产费用之间的相关关系分析中，产量是自变量，用来解释生产费用，所以产量为 x，生产费用为 y。

定义 8.5　在回归分析中，用来测量非自变量之外因素对因变量的影响，称为误差项，用 ξ 表示。

定义 8.6　描述因变量 y 如何依赖于自变量 x 和误差项 ξ 的方程，称为回归模型。

只涉及一个自变量的一元线性回归模型可表示为

$$y = \beta_0 + \beta_1 x + \xi \tag{8.2}$$

在上述一元线性回归模型中，y 是 x 的线性函数加上误差项 ξ。其中 β_0、β_1 称为模型的参数。

对于一元回归模型，有以下几个基本的假定。

(1) 误差项 ξ 服从正态分布，即 $\xi \sim N(0,\sigma^2)$，数学期望值为 0，方差为 σ^2。其中数学期望值为 0，是采用最小二乘法可以保证的；如果方差不为 σ^2，即不是定值时，在计量经济学中称之为异方差，要用相应的修正方法来使之满足。

(2) 误差项 ξ 与解释变量 x 无关，即 $\mathrm{cov}(x,\xi)=0$。上面介绍过，除了产量影响生产费用之外，还有其他的因素，只是产量占主要的因素，这时要求误差项 ξ 与解释变量 x 无关，否则会出现多重共线性。

(3) 在重复抽样中，解释变量 x 是固定的，即假定 x 是非随机的。

根据回归模型中的假定，ξ 的期望等于 0，因此 y 的期望值 $E(y)=\beta_0+\beta_1 x$。也就是说，y 的期望值是 x 的线性函数。

定义 8.7 描述因变量 y 的期望值如何依赖于自变量 x 的方程，称为回归方程。

一元线性回归方程的形式为

$$E(y) = \beta_0 + \beta_1 x \tag{8.3}$$

式中，β_0 为回归直线在 y 轴上的截距，是 $x=0$ 时 y 的期望值；β_1 为直线的斜率，它表示当 x 每变动一个单位时，y 平均变动值。

如果回归方程中的参数 β_0、β_1 已知，那对于一个给定的 x 值，利用 $E(y) = \beta_0 + \beta_1 x$ 就能计算出 y 的期望值。但由于总体回归参数 β_0、β_1 是未知的，因此必须利用样本数据去估计它们。

2. 样本回归方程

定义 8.8 根据样本数据求出的回归方程的估计，称为样本回归方程，又称估计的回归方程。

则一元线性回归、样本的回归方程可表示为

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i \tag{8.4}$$

式中，$\hat{\beta}_0$ 为样本回归方程的截距；$\hat{\beta}_1$ 为样本回归方程的斜率。表示为给定解释变量 x_i 值，得到被解释变量的估计值。

8.2.3 参数的最小二乘估计

要根据给出的样本点来计算出 $\hat{\beta}_0$、$\hat{\beta}_1$ 的数值，这里使用的是最小二乘法。

首先介绍一下什么是最小二乘法。

在初中几何作图中，我们学过如何根据若干个点来做一条直线。以例 8.1 的数据为例来回忆一下，如何做出这条直线的，即做线的原则是什么。

第一步：先根据样本点，在坐标轴上描点，得到散点图。

第二步：根据做线的原则，画出一条直线。

下面就要把这条直线求出来，即求样本回归方程 $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$，如图 8.4 所示。

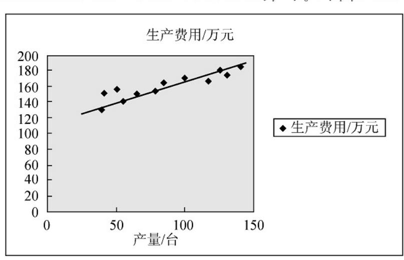


图 8.4 样本回归方程示意图

这里所说的最小二乘法就是根据这个做线的原则得出的。

做线的原则：让尽可能多的点落在直线上；如果落不到直线上的点，让其尽可能地在

直线上下近距离的波动。

让尽可能多的点落在直线上，落在直线上的点到直线的距离就是 0，是希望这样的点越多越好；不能落在直线上的点，让其在直线近距离地波动，也是希望点到直线的距离越小越好，也就是说，它希望得到的这条直线是让所有点到直线的距离和最小。这时出现一个关键词“最小”。

问题又出现，点到直线的距离和最小，那这里所说的距离是哪个距离？一个点到直线的距离有 3 种：竖直距离、垂直距离和水平距离，如图 8.5 所示。

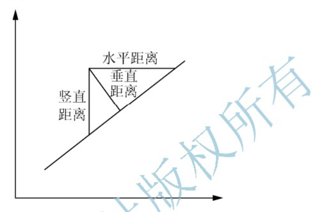


图 8.5　点到直线的 3 种距离示意图

由于在一元回归模型假定 x 是非随机的，因此取的是竖直距离。

由上面的分析可知，要所有点到直线的竖直距离和最小。每个点到直线竖直的距离表示 $y_i-\hat{y}_i$，如果点在直线上面，此竖直距离为正；如果点在直线下面，此竖直距离为负，要求的是距离和最小，而竖直距离有正有负，无法求出和最小，因此把每一个点的竖直距离进行平方再求和最小，即求每个点到直线竖直距离平方和最小，这里最后出现了两个关键词“平方”、“最小”，合起来就是最小平方法，即最小二乘法。

定义 8.9　使因变量的观察值 y_i 与估计值 $\hat{y}_i$ 之间的离差平方和达到最小来求得 $\hat{\beta}_0$ 和 $\hat{\beta}_0$ 的方法，称为最小二乘法，也称最小平方法。

最小二乘法求解参数估计值的步骤如下。

(1) 根据最小二乘法的定义可得

$$L=\sum(y_i-\hat{y}_i)^2=\sum(y_i-\hat{\beta}_0-\hat{\beta}_1x_i)^2$$

(2) 求 $\min L$，一般情况，分别对上式中的未知数求一阶偏导，令其式子为 0。

$$\frac{\partial L}{\partial\hat{\beta}_0}=\sum -2\times(y_i-\hat{\beta}_0-\hat{\beta}_1x_i)=0 \qquad ①$$

$$\frac{\partial L}{\partial\hat{\beta}_1}=\sum -2\times(y_i-\hat{\beta}_0-\hat{\beta}_1x_i)x_i=0 \qquad ②$$

(3) 式①整理：

$$\sum y_i-n\hat{\beta}_0-\hat{\beta}_1\sum x_i=0$$

两边同时除以 n，可以得到

$$\hat{\beta}_0=\overline{y}-\hat{\beta}_1\overline{x}$$

式②整理：

$$\sum x_i y_i - \hat{\beta}_0 \sum x_i - \hat{\beta}_1 \sum x_i^{\ 2} = 0$$

将 $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$ 和 $\sum x_i = n\bar{x}$ 代入上式中，整理得

$$\sum x_i y_i - (\bar{y} - \hat{\beta}_1 \bar{x}) \times n\bar{x} - \hat{\beta}_1 \sum x_i^{\ 2} = 0$$

$$\hat{\beta}_1 = \frac{\sum x_i y_i - n\overline{xy}}{\sum x_i^{\ 2} - n\bar{x}^2}$$

即有

$$\hat{\beta}_1 = \frac{\sum x_i y_i - n\overline{xy}}{\sum x_i^{\ 2} - n\bar{x}^2} \text{ 或者 } \hat{\beta}_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \tag{8.5}$$

【例 8.4】 沿用例 8.1 中的数据，计算产量 x 与生产费用 y 的样本回归方程。

解：最小二乘法的公式为

$$\begin{aligned}\hat{\beta}_1 &= \frac{\sum x_i y_i - n\overline{xy}}{\sum x_i^{\ 2} - n\bar{x}^2} \\ &= \frac{170\,094 - 12 \times 85.42 \times 160.08}{101\,835 - 12 \times 85.42^2} \\ &= \frac{6\,008.583}{14\,282.92} = 0.42\end{aligned}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 160.08 - 0.42 \times 85.42 = 124.20$$

即样本回归方程为 $\hat{y}_i = 124.20 + 0.42x_i$。

从最小二乘法的结果可以得出一个重要的性质，即 $(\bar{x}, \bar{y})$ 在样本回归方程上 $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$，因为有 $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$，整理即为 $\bar{y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{x}$。

可以利用 Excel 进行回归方程的计算，在本章 8.5 节案例中会详细介绍。

8.2.4 样本回归方程的评价

前面介绍了最小二乘法是使因变量的观察值 y_i 与估计值 $\hat{y}_i$ 之间的离差平方和达到最小来求得 $\hat{\beta}_0$ 和 $\hat{\beta}_0$ 的方法，最后得出样本回归方程 $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$，下面要对样本回归方程进行评价。这里所说的评价，就是对直线的拟合程度进行评价，可以想象，如果各观测数据的散点都落在直线上，那么这条直线就是对数据完全的拟合，同时代表各具点，x 得到的估计值与真实值 y 是没有误差的，即各离散点越紧密围绕直线，这条直线回归得越好，拟合得越好。

定义 8.10　回归直线与各观察点的接近程度，称为回归直线对数据的拟合优度。

评价的指标有很多，这里只介绍两个指标。

1．判定系数

判定系数是对样本回归方程拟合程度的一个度量，即围绕着每一个观察点，x 得到的估计值与真实值 y 的误差大小进行测量。

用最小二乘法估计出样本回归方程后，会有两组数据，分别为真实值 y 和 $\hat{y}$，即 $y_1, y_2, \cdots, y_n$ 和 $\hat{y}_1, \hat{y}_2, \cdots, \hat{y}_n$。

对于 $y_1, y_2, \cdots, y_n$ 这组数据，数据之间存在差异，差异大小用差异的平方和表示，称这种差异的大小为总平方和，记为 SST。总平方和的公式为

$$\text{SST} = \sum (y_i - \overline{y})^2 \tag{8.6}$$

总平方和是反映真实值的离散程度。而总平方和恰好又可以分解，如图 8.6 所示。

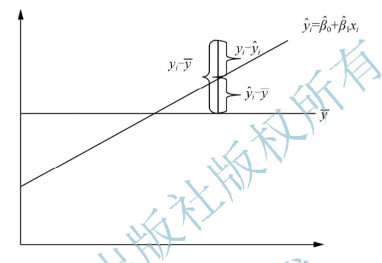


图 8.6　总平方和分解示意图

由图 8.6 可得

$$\text{SST} = \sum (y_i - \overline{y})^2 = \sum (y_i - \hat{y}_i)^2 + \sum (\hat{y}_i - \overline{y})^2$$

其中 $\sum (\hat{y}_i - \overline{y})^2$ 是反映 $\hat{y}_1, \hat{y}_2, \cdots, \hat{y}_n$ 这组数据的离散程度。因为有 $\overline{\hat{y}} = \overline{y}$，而 $\hat{y}_1, \hat{y}_2, \cdots, \hat{y}_n$ 是根据样本回归方程得出的，所以称为回归平方和，记为 SSR；而 $\sum (y_i - \hat{y}_i)^2$ 是表示各实际观测点与回归直线的残差 $y_i - \hat{y}_i$ 的平方和，称为残差平方和，记为 SSE。即有

$$\text{SST} = \text{SSR} + \text{SSE} \tag{8.7}$$

从图 8.6 可以直观地看出，样本回归方程拟合的好坏取决于 SSR 和 SST 的比例，如果全部点落在直线上，此时 SSE 为 0，而 SSR 与 SST 是相等的，也就是说样本数据 100%都落在直线上了，表示直线拟合越好；如果全部点都没落在直线上，即样本观察点 0%落在直线上，表示直线拟合不好。

定义 8.11　回归平方和占总平方和的比例，称为判定系数，记为 R^2。

$$R^2 = \frac{\text{SSR}}{\text{SST}} = \frac{\text{SST} - \text{SSE}}{\text{SST}} = 1 - \frac{\text{SSE}}{\text{SST}} \tag{8.8}$$

由式(8.8)可知判定系数 R^2 的取值范围为 $[0,1]$，当 $R^2 = 1$ 时，代表所有样本点都落在回归直线上，这时这条回归直线是完全拟合；当 $R^2 = 0$ 时，代表没有样本点落在回归直线上，这时回归直线拟合的是最差的。所以 R^2 越接近 1，回归直线的拟合越好。

【例 8.5】 沿用例 8.1 中的数据，用判定系数评价例 8.4 得到的产量 x 与生产费用 y 的样本回归方程。

解：根据数据可得 $\overline{y}=160$，所以有

$$\begin{aligned}\text{SST}&=\sum(y_i-\overline{y})^2\\&=(130-160)^2+\cdots+(185-160)^2\\&=2\,985\end{aligned}$$

在例 8.4 中，已经得出样本回归方程为 $\hat{y}_i=124.20+0.42x_i$，把样本点中产量 x_i 代入回归直线中，可得出 $\hat{y}_i$ 的数值。

$$\text{SSE}=\sum(y_i-\hat{y}_i)^2=457.211$$

其中有 $\text{SST}=\text{SSR}+\text{SSE}$，则 $\text{SSR}=2527.79$，所以有

$$R^2=\frac{\text{SSR}}{\text{SST}}=\frac{\text{SST}-\text{SSE}}{\text{SST}}=1-\frac{\text{SSE}}{\text{SST}}=84.68\%$$

即有 84.68% 个样本点落在直线上，说明直线拟合优度较好。

从判定系数的分析过程可以得知 $\text{SST}=\sum(y_i-\overline{y})^2$、$\text{SSE}=\sum(y_i-\hat{y}_i)^2$ 和 $\text{SSR}=\sum(\hat{y}_i-\overline{y})^2$，同时在进行回归分析时，对回归分析做了几条基本的假定，其中有一条是误差项 ξ 服从正态分布，即 $\xi\sim N(0,\sigma^2)$，既然误差项服从正态分布，可以推知因变量 y 也服从正态分布，所以有 $\overline{y}$ 服从正态分布，最后可以推出 $\text{SST}=\sum(y_i-\overline{y})^2$、$\text{SSE}=\sum(y_i-\hat{y}_i)^2$ 和 $\text{SSR}=\sum(\hat{y}_i-\overline{y})^2$ 三者都服从 χ^2 分布，且自由度分别为 $n-1$、$n-k-1$、k，其中 k 为解释变量的个数，一元线性回归解释变量个数为 1，即有

$$\begin{aligned}&\text{SST}\sim\chi^2(n-1)\\&\text{SSR}\sim\chi^2(k)\\&\text{SSE}\sim\chi^2(n-k-1)\end{aligned}\tag{8.9}$$

2. 估计标准误差

残差平方和 $\text{SSE}=\sum(y_i-\hat{y}_i)^2$，表示实际观测值 y_i 与样本回归方程的估计值 $\hat{y}_i$ 之间的差异程度，也可以用来度量各实际观测值在直线周围的散布状况，这个量就是估计标准误差。

定义 8.12　残差均方的平方根，称为估计量的标准差，或称为标准误差，用 s_e 表示。

实质上，估计标准误差是对误差项 ξ 的标准差 σ 的估计。

估计标准误差的计算公式：

$$s_e=\sqrt{\text{MSE}}=\sqrt{\frac{\text{SSE}}{n-2}}\tag{8.10}$$

估计标准误差是反映实际观测值 y_i 与样本回归方程的估计值 $\hat{y}_i$ 之间的差异的大小，当 $s_e=0$ 时，代表各观测点全部都落在回归直线上，这条直线拟合的越好，所以说 s_e 越接近 0 时，样本回归方程拟合的越好。

【例 8.6】沿用例 8.1 中的数据，用估计标准误差评价例 8.4 得到的产量 x 与生产费用 y 的样本回归方程。

解：根据题意有

$$s_e=\sqrt{\text{MSE}}=\sqrt{\frac{\text{SSE}}{n-2}}$$

其中 $\text{SSE}=\sum(y_i-\hat{y}_i)^2=457.21$，$n-2=10$，所以有

$$s_e=\sqrt{\text{MSE}}=\sqrt{\frac{457.21}{10}}=6.76$$

8.2.5 一元线性回归方程的统计检验

从前面的介绍中，得到了样本的回归方程，但我们的目的不是要样本的回归方程，而是要总体的回归方程，这时利用推断统计来检验样本的回归方程是否能真实地反映解释变量 x 与被解释变量 y 的关系。

回归方程的统计检验主要包括两个方面的内容：一是线性关系的检验；二是回归系数的检验。

1. 线性关系的检验

线性关系的检验主要是检验解释变量 x 与被解释变量 y 之间的线性关系是否显著，即两者的线性模型 $y=\beta_0+\beta_1x+\xi$ 是否成立。

(1) 提出原假设 H_0 和备择假设 H_1。

$H_0:\beta_1=0$　线性关系不显著

$H_1:\beta_1\neq0$　线性关系显著

(2) 构造检验的统计量，并计算其值。

检验的统计量构造方法采用方差分析的方法来构造的，即以回归平方和(SSR)和残差平方和(SSE)为基础。

由上面的分析可得

$$\text{SST}\sim\chi^2(n-1)$$
$$\text{SSR}\sim\chi^2(k)$$
$$\text{SSE}\sim\chi^2(n-k-1)$$

其中 k 为解释变量的个数，在一元线性回归方程中，解释变量只有一个，即 $k=1$。所以检验的统计量为

$$F=\frac{\text{SSR}/1}{\text{SSE}/(n-2)}\sim F(1,n-2) \tag{8.11}$$

(3) 根据给出的显著性水平 α，确定拒绝原假设的区域。

$$(F_\alpha(1,n-2),+\infty)$$

(4) 统计决策。

当 $F>F_\alpha(1,n-2)$ 时，检验统计量落在拒绝原假设区域内，所以拒绝原假设，两变量的线性关系显著。

【例 8.7】沿用例 8.1 中的数据，检验产量 x 与生产费用 y 的线性关系是否显著。($\alpha=0.05$)

解：

(1) 提出原假设 H_0 和备择假设 H_1。

$H_0:\beta_1=0$　线性关系不显著

$H_1:\beta_1\neq0$　线性关系显著

(2) 构造检验的统计量，并计算其值。

$$F=\frac{\mathrm{SSR}/1}{\mathrm{SSE}/(n-2)}\sim F(1,n-2)$$

其中 $\mathrm{SSR}=2\,527.79$， $\mathrm{SSE}=457.21$，所以有

$$F=\frac{2\,527.79/1}{457.21/10}=55.287\,3$$

(3) $\alpha=0.05$，确定拒绝原假设的区域。

$$(F_{0.05}(1,10),+\infty)$$

查 F 分布表得 $F_{0.05}(1,10)=4.964\,603$。

(4) 统计决策。

$F=55.287\,3>F_{0.05}(1,10)=4.964\,603$，检验统计量落在拒绝原假设区域内，所以拒绝原假设，两变量的线性关系显著。

2. 回归系数的检验

在线性模型 $y=\beta_0+\beta_1x+\xi$ 中，回归系数有 β_0 和 β_1，所以要分别对这两个回归系数进行检验。

1) 回归系数 β_1 的检验

(1) 提出原假设 H_0 和备择假设 H_1。

$$H_0:\beta_1=0;\quad H_1:\beta_1\neq 0$$

(2) 构造检验的统计量，并计算其值。

回归系数 β_1 检验的统计量应从其样本入手，即 $\hat{\beta}_1$。根据最小二乘法有

$$\begin{aligned}\hat{\beta}_1&=\frac{\sum(x_i-\overline{x})(y_i-\overline{y})}{\sum(x_i-\overline{x})^2}\\&=\frac{\sum[(x_i-\overline{x})y_i-(x_i-\overline{x})\overline{y}]}{\sum(x_i-\overline{x})^2}\\&=\frac{\sum(x_i-\overline{x})y_i-\sum(x_i-\overline{x})\overline{y}}{\sum(x_i-\overline{x})^2}\\&=\frac{\sum(x_i-\overline{x})y_i-\overline{y}\sum(x_i-\overline{x})}{\sum(x_i-\overline{x})^2}\end{aligned}$$

其中 $\sum(x_i-\overline{x})=0$，所以有

$$\hat{\beta}_1=\frac{\sum(x_i-\overline{x})y_i}{\sum(x_i-\overline{x})^2}$$

即 $\hat{\beta}_1$ 是 y_i 的线性组合，而 y_i 服从正态分布，有 $y_i\sim N(\alpha+\beta x_i,\sigma^2)$，所以有 $\hat{\beta}_1$ 也服从正态分布。接着要计算其数学期望值和方差，计算过程如下。

令 $C_i=\dfrac{(x_i-\overline{x})}{\sum(x_i-\overline{x})^2}$，有 $\hat{\beta}_1=\sum C_iy_i$。 C_i 具有以下性质。

① $\sum C_i=0$。

证明：$\sum C_i = \sum \frac{(x_i - \bar{x})}{\sum (x_i - \bar{x})^2} = 0$

② $\sum C_i (x_i - \bar{x}) = 1$。

证明：
$$\sum C_i (x_i - \bar{x}) = \sum \frac{(x_i - \bar{x})}{\sum (x_i - \bar{x})^2}(x_i - \bar{x})$$
$$= \sum \frac{(x_i - \bar{x})^2}{\sum (x_i - \bar{x})^2}$$
$$= \frac{\sum (x_i - \bar{x})^2}{\sum (x_i - \bar{x})^2} = 1$$

③ $\sum C_i x_i = 1$。

证明：有 $\sum C_i (x_i - \bar{x}) = 1$

左边 $= \sum (C_i x_i - \bar{x} C_i) = \sum C_i x_i - \sum \bar{x} C_i = \sum C_i x_i - \bar{x} \sum C_i$

其中 $\sum C_i = 0$，所以有 $\sum C_i x_i = 1$。

④ $\sum C_i^2 = \frac{1}{\sum (x_i - \bar{x})^2}$。

证明：
$$\sum C_i^2 = \sum \left(\frac{(x_i - \bar{x})}{\sum (x_i - \bar{x})^2}\right)^2 = \sum \frac{(x_i - \bar{x})^2}{\left(\sum (x_i - \bar{x})^2\right)^2}$$
$$= \frac{\sum (x_i - \bar{x})^2}{\left(\sum (x_i - \bar{x})^2\right)^2}$$
$$= \frac{1}{\sum (x_i - \bar{x})^2}$$

有了上面 C_i 的性质后，下面来计算一下 $\hat{\beta}_1$ 的数学期望值和方差。

$$E(\hat{\beta}_1) = E\left(\sum C_i y_i\right) = \sum C_i E(y_i) = \sum C_i (\alpha + \beta x_i) = \alpha \sum C_i + \beta \sum C_i x_i$$

其中 $\sum C_i = 0$，$\sum C_i x_i = 1$，则有

$$E(\hat{\beta}_1) = \beta_1$$

$D(\hat{\beta}_1) = D\left(\sum C_i y_i\right)$，因为 y_i 之间是相互独立的，所以有

$$D(\hat{\beta}_1) = \sum C_i^2 D(y_i)$$

其中 $D(y_i) = \sigma^2$，则有

$$D(\hat{\beta}_1) = \sigma^2 \sum C_i^2 = \frac{1}{\sum (x_i - \bar{x})^2} \sigma^2$$

即有

$$\hat{\beta}_1 \sim N\left(\beta_1, \frac{1}{\sum (x_i - \bar{x})^2} \sigma^2\right) \tag{8.12}$$

有了这个分布，下面开始构造检验的统计量。

$$\hat{\beta}_1 \sim N\left(\beta_1, \frac{1}{\sum (x_i - \bar{x})^2} \sigma^2\right)$$

$$Z=\frac{\hat{\beta}_1-\beta_1}{\sqrt{\frac{1}{\sum(x_i-\bar{x})^2}\sigma^2}}\sim N(0,1)$$

因为σ^2未知，而前面已介绍估计标准误差是其估计值，而估计标准误差是残差均方，所以有

$$t=\frac{\hat{\beta}_1-\beta_1}{\sqrt{\frac{1}{\sum(x_i-\bar{x})^2}s_e^2}}\sim t(n-2) \tag{8.13}$$

把原假设$H_0:\beta_1=0$代入此式，有

$$t=\frac{\hat{\beta}_1}{s_{\hat{\beta}_1}}\sim t(n-2) \tag{8.14}$$

式中，$s_{\hat{\beta}_1}=\sqrt{\frac{1}{\sum(x_i-\bar{x})^2}s_e^2}$，是$\hat{\beta}_1$的标准误差。

(3) 根据给出的显著性水平α，确定拒绝原假设的区域。

此检验为双侧检验，其拒绝原假设的区域为

$$(-\infty,-t_{\alpha/2}(n-2))\cup(t_{\alpha/2}(n-2),+\infty)$$

(4) 统计决策。

当$|t|>t_{\alpha/2}(n-2)$时，检验统计量落在拒绝原假设区域内，所以拒绝原假设，回归系数β_1显著。

2) 回归系数β_0的检验

(1) 提出原假设H_0和备择假设H_1。

$$H_0:\beta_0=0;\quad H_1:\beta_0\neq 0$$

(2) 构造检验的统计量，并计算其值。

回归系数β_0检验的统计量应从其样本入手，即$\hat{\beta}_0$。根据最小二乘法有

$$\hat{\beta}_0=\bar{y}-\hat{\beta}_1\bar{x}=\frac{1}{n}(y_1+y_2+\cdots+y_n)-\hat{\beta}_1\bar{x}$$

其中$\hat{\beta}_1=\sum C_i y_i$，所以有

$$\begin{aligned}\hat{\beta}_0&=\frac{1}{n}(y_1+y_2+\cdots+y_n)-\bar{x}\sum C_i y_i\\&=\sum(\frac{1}{n}-\bar{x}C_i)y_i\end{aligned}$$

同样$\hat{\beta}_0$是y_i的线性组合，而y_i服从正态分布，有$y_i\sim N(\alpha+\beta x_i,\sigma^2)$，所以有$\hat{\beta}_0$也服从正态分布。接着要计算其数学期望值和方差，计算过程如下。

$$\begin{aligned}E(\hat{\beta}_0)&=\sum(\frac{1}{n}-\bar{x}C_i)E(y_i)\\&=\sum(\frac{1}{n}-\bar{x}C_i)(\alpha+\beta x_i)\end{aligned}$$

$$= \alpha \sum (\frac{1}{n} - \bar{x}C_i) + \beta \sum (\frac{1}{n} - \bar{x}C_i)x_i$$
$$= \alpha(1 - \bar{x}\sum C_i) + \beta(\bar{x} - \bar{x}\sum C_i x_i)$$

其中$\sum C_i = 0$，$\sum C_i x_i = 1$，有

$$E(\hat{\beta}_0) = \beta_0$$

$$D(\hat{\beta}_0) = D(\sum (\frac{1}{n} - \bar{x}C_i)y_i) = \sum (\frac{1}{n} - \bar{x}C_i)^2 D(y_i)$$
$$= \sigma^2 \sum (\frac{1}{n} - \bar{x}C_i)^2$$
$$= \sigma^2 \sum (\frac{1}{n^2} - 2\bar{x}C_i + \bar{x}^2 C_i^2)$$
$$= \sigma^2 (\frac{1}{n} - 2\bar{x}\sum C_i + \bar{x}^2 \sum C_i^2)$$

其中$\sum C_i = 0$，$\sum C_i^2 = \dfrac{1}{\sum (x_i - \bar{x})^2}$，有

$$D(\hat{\beta}_0) = \left(\frac{1}{n} + \frac{\bar{x}^2}{\sum (x_i - \bar{x})^2}\right)\sigma^2$$

即有

$$\hat{\beta}_0 \sim N\left(\beta_0, \left(\frac{1}{n} + \frac{\bar{x}^2}{\sum (x_i - \bar{x})^2}\right)\sigma^2\right)$$

$$Z = \frac{\hat{\beta}_0 - \beta_0}{\sqrt{\left(\frac{1}{n} + \frac{\bar{x}^2}{\sum (x_i - \bar{x})^2}\right)\sigma^2}} \sim N(0,1)$$

同样因为σ^2未知，用估计标准误差代替，所以有

$$t = \frac{\hat{\beta}_0 - \beta_0}{\sqrt{\left(\frac{1}{n} + \frac{\bar{x}^2}{\sum (x_i - \bar{x})^2}\right)s_e^{\ 2}}} \sim t(n-2) \tag{8.15}$$

其中把原假设$H_0: \beta_0 = 0$代入此式，有

$$t = \frac{\hat{\beta}_0}{s_{\hat{\beta}_0}} \sim t(n-2) \tag{8.16}$$

其中$s_{\hat{\beta}_0} = \sqrt{\left(\frac{1}{n} + \frac{\bar{x}^2}{\sum (x_i - \bar{x})^2}\right)s_e^2}$，是$\hat{\beta}_0$的标准误差。

(3) 根据给出的显著性水平α，确定拒绝原假设的区域。

此检验为双侧检验，其拒绝原假设的区域为

$$(-\infty, -t_{\alpha/2}(n-2)) \cup (t_{\alpha/2}(n-2), +\infty)$$

(4) 统计决策。

当$|t| > t_{\alpha/2}(n-2)$时，检验统计量落在拒绝原假设区域内，所以拒绝原假设，回归系数β_0显著。

【例 8.8】 沿用例 8.1 中的数据，检验回归模型的回归系数是否显著。($\alpha = 0.05$)

解：首先检验回归系数β_1。

(1) 提出原假设H_0和备择假设H_1。

$$H_0: \beta_1 = 0；\quad H_1: \beta_1 \neq 0$$

(2) 构造检验的统计量，并计算其值

$$t = \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\dfrac{1}{\sum (x_i - \overline{x})^2} s_e^2}} \sim t(n-2)$$

把原假设$H_0: \beta_1 = 0$代入此式，有

$$t = \frac{\hat{\beta}_1}{s_{\hat{\beta}_1}} = \frac{0.42}{\sqrt{\dfrac{457.21}{14\,282.92}}} = \frac{0.42}{\sqrt{0.003\,201}} = 7.423\,349$$

(3) $\alpha = 0.05$，确定拒绝原假设的区域。

$$(-\infty, -t_{0.025}(10)) \cup (t_{0.025}(10), +\infty)$$

查 t 分布表可得$t_{0.025}(10) = 2.633\,767$。

(4) 统计决策。

因为$t = 7.423\,349 > t_{0.025}(10) = 2.633\,767$，所以拒绝原假设，即回归系数$\beta_1$显著。

最后检验回归系数β_0。

(1) 提出原假设H_0和备择假设H_1。

$$H_0: \beta_0 = 0；\quad H_1: \beta_0 \neq 0$$

(2) 构造检验的统计量，并计算其值。

$$t = \frac{\hat{\beta}_0 - \beta_0}{\sqrt{\left(\dfrac{1}{n} + \dfrac{\overline{x}^2}{\sum (x_i - \overline{x})^2}\right) s_e^2}} \sim t(n-2)$$

把原假设$H_0: \beta_0 = 0$代入此式，有

$$t = \frac{\hat{\beta}_0}{s_{\hat{\beta}_0}} = \frac{124.20}{\sqrt{\left(\dfrac{1}{12} + \dfrac{7\,296.007}{14\,282.92}\right) \times 45.721}} = \frac{124.20}{\sqrt{27.165\,31}} = \frac{124.20}{5.212\,035} = 23.829\,46$$

(3) $\alpha = 0.05$，确定拒绝原假设的区域。

$$(-\infty, -t_{0.025}(10)) \cup (t_{0.025}(10), +\infty)$$

查 t 分布表可得$t_{0.025}(10) = 2.633\,767$。

(4) 统计决策。

因为$t = 23.829\,46 > t_{0.025}(10) = 2.633\,767$，所以拒绝原假设，即回归系数$\beta_0$显著。

8.3 一元线性回归模型的预测

回归分析的主要目的是根据所建立的估计回归方程进行经济预测和经济控制。所谓预测，是指通过自变量 x 的取值来预测因变量 y 的取值。控制是指通过因变量 y 值，求出自变量 x 的值。这里主要介绍根据估计方程进行估计和预测的方法，主要包括点估计和区间估计。

8.3.1 点估计

定义 8.13 利用估计的回归方程，对于 x 的一个特定值 x_0，求出 y 的一个估计值就是点估计。

点估计分两种：①平均值的点估计；②个别值的点估计。

定义 8.14 利用估计的回归方程，对于 x 的一个特定值 x_0，求出 y 的平均值的一个估计值 $E(y_0)$，称为平均值的点估计。

【例 8.9】 沿用例 8.1 中的数据，利用估计的回归方程，对于一个特定值 $x_0 = 90$，求出 y_0 平均值的点估计。

解：平均值的一个估计值为

$$E(y_0) = 124.20 + 0.42 \times 90 = 162$$

定义 8.15 利用估计的回归方程，对于 x 的一个特定值 x_0，求出 y 的一个个别值的估计值 $\hat{y}_0$，称为个别值的点估计。

【例 8.10】 沿用例 8.1 中的数据，利用估计的回归方程，对于一个特定值 $x_0 = 90$，求出 y_0 个别值的点估计。

解：y_0 个别值的点估计为

$$\hat{y}_0 = 124.20 + 0.42 \times 90 = 162$$

从上面的两个例题中可以看出，在点估计的条件下，对于同一个 x_0，平均值的点估计和个别值的点估计的结果是一样的。但是，两者在区间中则有所不同。

8.3.2 区间估计

定义 8.16 利用估计的回归方程，对于 x 的一个特定值 x_0，求出 y 的一个估计值的区间就是区间估计。

区间估计也有两种类型：置信区间估计和预测区间估计。

定义 8.17 对 x 的一个给定值 x_0，求出 y 的平均值的估计区间，这一区间称为置信区间。

定义 8.18 对 x 的一个给定值 x_0，求出 y 的一个个别值的估计区间，这一区间称为预测区间。

1. 置信区间估计

置信区间估计是对一个给定值 x_0，求出 y 的平均值的估计区间，即求出 $E(y_0)$ 的区间

范围。因为 $y_0 = \hat{y}_0 + \xi_0$，其中 $E(\xi_0) = 0$，所以有 $E(y_0) = E(\hat{y}_0)$，即 $\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$ 为 $E(y_0)$ 的估计值。

已知 $\hat{\beta}_0$ 和 $\hat{\beta}_1$ 都服从正态分布，而 x_0 是给定的一个值，所以 $\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$ 也服从正态分布。既然服从正态分布，那么下面就来计算其方差。

$$
\begin{aligned}
D(\hat{y}_0) &= D(\hat{\beta}_0 + \hat{\beta}_1 x_0) \\
&= D(\hat{\beta}_0) + D(\hat{\beta}_1 x_0) + 2\operatorname{cov}(\hat{\beta}_0, \hat{\beta}_1 x_0) \\
&= \left(\frac{1}{n} + \frac{\overline{x}^2}{\sum (x_i - \overline{x})^2}\right)\sigma^2 + x_0^2 \times \frac{1}{\sum (x_i - \overline{x})^2}\sigma^2 + 2x_0 \operatorname{cov}(\hat{\beta}_0, \hat{\beta}_1)
\end{aligned}
$$

其中
$$
\begin{aligned}
\operatorname{cov}(\hat{\beta}_0,\ \hat{\beta}_1) &= \operatorname{cov}(\overline{y} - \hat{\beta}_1 \overline{x}\ ,\hat{\beta}_1) \\
&= \operatorname{cov}(\overline{y}, \hat{\beta}_1 \overline{x}) - \operatorname{cov}(\hat{\beta}_1 \overline{x},\ \hat{\beta}_1) \\
&= 0 - \overline{x}\operatorname{cov}(\hat{\beta}_1,\ \hat{\beta}_1) = -\overline{x}D(\hat{\beta}_1) \\
&= -\overline{x}\frac{1}{\sum (x_i - \overline{x})^2}\sigma^2
\end{aligned}
$$

$$
D(\hat{y}_0) = \left(\frac{1}{n} + \frac{\overline{x}^2 + x_0^2 - 2x_0\overline{x}}{\sum (x_i - \overline{x})^2}\right)\sigma^2 = \left(\frac{1}{n} + \frac{(x_0 - \overline{x})^2}{\sum (x_i - \overline{x})^2}\right)\sigma^2
$$

由于总体误差项 ξ 的方差 σ^2 是未知的，用估计的标准误差来代替，但此时是服从 t 分布的，有

$$
s_{\hat{y}_0} = s_e \sqrt{\left(\frac{1}{n} + \frac{(x_0 - \overline{x})^2}{\sum (x_i - \overline{x})^2}\right)} \tag{8.17}
$$

对于给定值 x_0，求出 y 的平均值的估计区间为

$$
\hat{y}_0 \pm t_{\alpha/2}(n-2) s_e \sqrt{\left(\frac{1}{n} + \frac{(x_0 - \overline{x})^2}{\sum (x_i - \overline{x})^2}\right)} \tag{8.18}
$$

【例 8.11】 沿用例 8.1 中的数据，求 12 家企业生产费用 95%的置信区间。($\alpha = 0.05$)

解：根据题意，置信区间为

$$
\hat{y}_0 \pm t_{\alpha/2}(n-2) s_e \sqrt{\left(\frac{1}{n} + \frac{(x_0 - \overline{x})^2}{\sum (x_i - \overline{x})^2}\right)}
$$

其中，$\hat{y}_0 = 162$，$t_{0.025}(10) = 2.633\,767$，$s_e = 6.76$，所以有

$$
\begin{aligned}
&162 \pm 2.633\,767 \times 6.76 \times \sqrt{\left(\frac{1}{12} + \frac{(90 - 85.42)^2}{14\,282.92}\right)} \\
&= 162 \pm 2.633\,767 \times 6.76 \times 0.291\,208 \\
&= 162 \pm 5.184\,74
\end{aligned}
$$

即置信区间为 $(156.815\,3, 167.184\,7)$。

2. 预测区间估计

预测区间是指对 x 的一个给定值 x_0，求出 y 的一个个别值的估计区间，即求出 $\hat{y}_0 + \xi_0$ 的区间范围。

从上面的分析中可知 $\hat{y}_0=\hat{\beta}_0+\hat{\beta}_1x_0$ 为 $E(y_0)$ 的估计值，而预测区间估计是求 $\hat{y}_0+\xi_0$ 的范围，比置信区间多了一个误差项 ξ_0，即

$$D(\hat{y}_0+\xi_0)=D(\hat{y}_0)+D(\xi_0)=\left(\frac{1}{n}+\frac{(x_0-\overline{x})^2}{\sum(x_i-\overline{x})^2}\right)\sigma^2+\sigma^2=\left(1+\frac{1}{n}+\frac{(x_0-\overline{x})^2}{\sum(x_i-\overline{x})^2}\right)\sigma^2$$

同样用估计的标准误差来代替方差 σ^2，是服从 t 分布的，有

$$s_{y_0}=s_e\sqrt{\left(1+\frac{1}{n}+\frac{(x_0-\overline{x})^2}{\sum(x_i-\overline{x})^2}\right)} \tag{8.19}$$

对于给定值 x_0，求出 y 的平均值的估计区间为

$$\hat{y}_0\pm t_{\alpha/2}(n-2)s_e\sqrt{\left(1+\frac{1}{n}+\frac{(x_0-\overline{x})^2}{\sum(x_i-\overline{x})^2}\right)} \tag{8.20}$$

【例 8.12】 沿用例 8.1 中的数据，求 12 家企业生产费用 95%的预测区间。

解：根据题意，置信区间为

$$\hat{y}_0\pm t_{\alpha/2}(n-2)s_e\sqrt{\left(1+\frac{1}{n}+\frac{(x_0-\overline{x})^2}{\sum(x_i-\overline{x})^2}\right)}$$

其中，$\hat{y}_0=162$，$t_{0.025}(10)=2.633\,767$，$s_e=6.76$，所以有

$$\begin{aligned}&162\pm2.633\,767\times6.76\times\sqrt{\left(1+\frac{1}{12}+\frac{(90-85.42)^2}{14\,282.92}\right)}\\&=162\pm2.633\,767\times6.76\times1.041\,538\\&=162\pm18.543\,82\end{aligned}$$

即置信区间为 $(143.456\,2,180.543\,8)$。

8.4　多元线性回归分析

在实际问题中，影响因变量的因素往往有很多，这种一个因变量与多个自变量的回归问题就是多元回归。当因变量与各自变量之间为线性关系时，称为多元线性回归。多元线性回归分析的原理与一元线性回归分析的原理基本相同，但在计算上要复杂得多，因而需要借助计算机来完成。这里只介绍多元与一元不同的内容。

8.4.1　多元线性回归模型的含义

多元线性回归相对于一元线性回归不同的是自变量的个数不是一个，而是多个，即两个及两个以上。

定义 8.19　描述因变量 y 如何依赖于自变量 $x_1,x_2,\cdots,x_k$ 和误差项 ξ 的方程，称为多元线性回归模型。

多元线性回归模型的一般形式可表示为

$$y=\beta_0+\beta_1x_1+\beta_2x_2+\cdots+\beta_kx_k+\xi \tag{8.21}$$

式中，$\beta_0,\beta_1,\beta_2,\cdots,\beta_k$ 为模型参数；ξ 为误差项。与一元线性回归类似，我们对误差项 ξ 有同样的基本假定。

定义 8.20　描述 y 的期望值如何依赖于 $x_1, x_2, \cdots, x_k$ 的方程，称为多元线性回归方程，即

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k \tag{8.22}$$

由于回归方程中的参数 $\beta_0, \beta_1, \beta_2, \cdots, \beta_k$ 是未知的，因而需要利用样本数据去估计它们。在用样本统计量 $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \cdots, \hat{\beta}_k$ 估计回归方程中的未知参数 $\beta_0, \beta_1, \beta_2, \cdots, \beta_k$ 时，就得到了估计的多元回归方程。

定义 8.21　根据样本数据得到的多元线性回归方程的估计，称为估计的多元线性回归方程。

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_k x_k \tag{8.23}$$

8.4.2　最小二乘法

回归方程中的 $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \cdots, \hat{\beta}_k$ 仍然是根据最小二乘法求得：

$$L = \sum (y_i - \hat{y}_i)^2 = \sum (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_1 - \hat{\beta}_2 x_2 - \cdots - \hat{\beta}_k x_k)^2$$

$$\begin{cases} \dfrac{\partial L}{\partial \hat{\beta}_0} = 0 \\ \dfrac{\partial L}{\partial \hat{\beta}_i} = 0 \end{cases} \quad i = 1, 2, \cdots, k \tag{8.24}$$

要从上面方程组中解出参数的估计值，需要借助计算机来完成。

8.4.3　样本回归方程的评价

类似于一元回归，对于多元线性回归方程，需要用指标评价多元回归方程的拟合优度。这里采用的指标有两个：一是修正多重判定系数；二是估计标准误差。其中估计标准误差同一元的计算公式相同，而多重判定系数是对一元的判定系数进行了修正。这里只介绍修正的多重判定系数。

在一元回归中曾介绍了因变量变差平方和的分解，这一点同样适用于多元回归中因变量变差平方和的分解，即

$$\text{SST} = \text{SSE} + \text{SSR}$$

式中，$\text{SST} = \sum (y_i - \bar{y})^2$ 为总平方和；$\text{SSR} = \sum (\hat{y}_i - \bar{y})^2$ 为回归平方和；$\text{SSE} = \sum (y_i - \hat{y}_i)^2$ 为残差平方和。

定义 8.22　在多元回归中，回归平方和占总平方和的比例，称为多重判定系数。其计算公式为

$$R^2 = \frac{\text{SSR}}{\text{SSE}} = 1 - \frac{\text{SSE}}{\text{SST}}$$

由于自变量增加时，预测误差会变得比较小，从而减少残差平方和 SSE；由于回归平方和 SSR=SST−SSE，因此当 SSE 变小时，SSR 就会变大，进而使 R^2 变大。因此，如果模型中增加了一个自变量，那么即使这个自变量在统计上并不显著，R^2 也会变大。为避免因增加自变量而高估 R^2，统计学家提出用样本量 n 和自变量个数 k 去修正 R^2，以便计算出修正的多重判定系数。

定义 8.23　用模型中自变量的个数和样本量进行调整的多重判定系数，称为修正的多重判定系数，记为 R'^2。其计算公式为

$$R'^2 = 1 - \left(1 - R^2\right) \times \frac{n-1}{n-k-1} \tag{8.25}$$

8.4.4 显著性检验

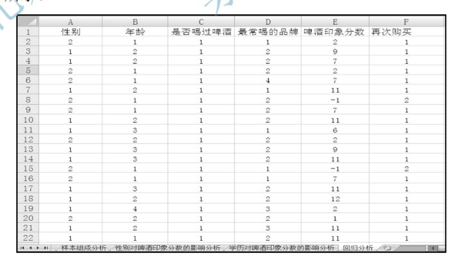
在一元线性回归中，线性关系的检验(F 检验)与回归系数的检验(t 检验)是等价的，这一点很容易理解。因为一元线性回归只有一个自变量。但在多元回归中，这两种检验不再等价。线性关系检验主要是检验因变量与多个自变量的线性关系是否显著，在 k 个自变量中，只要有一个自变量与因变量的线性关系显著，F 检验就能通过，但这不一定意味着每个自变量与因变量的关系都显著。回归系数检验则是对每个回归系数进行单独的检验，它主要用于检验每个自变量对因变量的影响是否显著。如果某个自变量没有通过检验，就意味着这个自变量对因变量的影响不显著，也就没有必要将这个自变量放进回归模型了。这部分内容主要通过案例来介绍。

8.5 案例分析：啤酒市场的调查与分析及 Excel 上机应用——啤酒销售量预测

回归分析常用的方法是向后消去法，向后消去法是指先纳入所有的变量，逐一消除没有预测效果的项目。这里以性别、年龄、是否喝过啤酒、何种品牌、啤酒印象为解释变量，是否购买为被解释变量，建立一个多元回归分析，以向后消去法最后得出一个有效的回归预测模型。

1. 建立一个五元线性回归预测模型

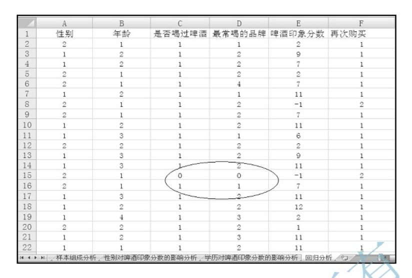
建立一个新的工作表，命名为“回归分析”，并且把资料库中的“性别”“年龄”“是否喝过啤酒”“最常喝的品牌”“啤酒印象分数”和“再次购买”数据复制到回归分析工作表中，如图 8.7 所示。



	A	B	C	D	E	F
1	性别	年龄	是否喝过啤酒	最常喝的品牌	啤酒印象分数	再次购买
2	2	1	1	1	2	1
3	1	2	1	2	9	1
4	1	2	1	2	7	1
5	2	1	1	2	2	1
6	2	1	1	4	7	1
7	1	2	1	1	11	1
8	2	1	1	2	-1	2
9	2	1	1	2	7	1
10	1	2	1	2	11	1
11	1	3	1	1	6	1
12	2	2	1	2	2	1
13	1	3	1	2	9	1
14	1	3	1	2	11	1
15	2	1	1	1	-1	2
16	2	1	1	1	7	1
17	1	3	1	2	11	1
18	1	2	1	2	12	1
19	1	4	1	3	2	1
20	2	2	1	2	1	1
21	1	2	1	3	11	1
22	1	1	1	2	11	1

图 8.7 “回归分析”工作表

为了研究变量“是否喝过啤酒”是否影响“再次购买”，将 C15 单元格和 D15 单元格的数值改为 0，如图 8.8 所示。



	A	B	C	D	E	F
1	性别	年龄	是否喝过啤酒	最常喝的品牌	啤酒印象分数	再次购买
2	2	1	1	1	2	1
3	1	2	1	2	9	1
4	1	2	1	2	7	1
5	2	1	1	2	2	1
6	2	1	1	4	7	1
7	1	2	1	1	11	1
8	2	1	1	2	-1	2
9	2	1	1	2	7	1
10	1	2	1	2	11	1
11	1	3	1	1	6	1
12	2	2	1	2	2	1
13	1	3	1	2	9	1
14	1	3	1	2	11	1
15	2	1	0	0	-1	2
16	2	1	1	1	7	1
17	1	3	1	2	11	1
18	1	2	1	2	12	1
19	1	4	1	3	2	1
20	2	2	1	2	1	1
21	1	2	1	3	11	1
22	1	1	1	2	11	1

样本组成分析 性别对啤酒印象分数的影响分析 学历对啤酒印象分数的影响分析 回归分析

图 8.8　将 C15 单元格和 D15 单元格的数值改为 0

建立五元线性回归预测模型步骤如下。

第一步：在回归分析工作表的界面，单击“数据”→“分析”→“数据分析”按钮，弹出“数据分析”对话框，在“分析工具”列表中选择“回归”选项，如图 8.9 所示。

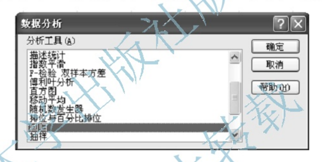


图 8.9　“数据分析”对话框

第二步：单击“确定”按钮，弹出“回归”对话框，在“Y 值输入区域”文本框中输入“\$F\$1:\$F\$31”，在“X 值输入区域”文本框中输入“\$A\$1:\$E\$31”，选择“标志”和“置信度”复选框，其中的置信度默认为“95%“，选中“输出区域”单选按钮，并在其文本框中输入输出结果“\$H\$9”，如图 8.10 所示。



图 8.10　“回归”对话框 1

第三步：单击“确定”按钮，出现如图 8.11 所示的分析结果。

SUMMARY OUTPUT									
回归统计									
Multiple R	0.778548235								
R Square	0.606137354								
Adjusted R Sc	0.524082636								
标准误差	0.175025031								
观测值	30								
方差分析									
	df	SS	MS	F	Significance F				
回归分析	5	1.131456394	0.226291279	7.386989665	0.000256788				
残差	24	0.735210273	0.030633761						
总计	29	1.866666667							
	Coefficients	标准误差	t Stat	P-value	Lower 95%	Upper 95%	下限 95.0%	上限 95.0%	
Intercept	2.356865098	0.325585243	7.238857263	1.76673E 07	1.684890189	3.028840007	1.684890189	3.02884001	
性别	-0.156909927	0.125228882	-1.252985137	0.222281961	-0.415369634	0.101549779	-0.41536963	0.10154978	
年龄	-0.075416303	0.051876464	-1.453767221	0.158963656	-0.182484062	0.031651456	-0.18248406	0.03165146	
是否喝过啤酒	-0.720147878	0.216155111	-3.331625489	0.002787937	-1.166270097	-0.27402566	-1.1662701	-0.2740257	
最常喝的品牌	-0.007531198	0.041996814	-0.179327834	0.859185671	-0.094208362	0.079145966	-0.09420836	0.07914597	
啤酒印象分数	-0.03237106	0.014017536	-2.309325915	0.029844544	-0.061301832	-0.00344029	-0.06130183	-0.0034403	

图 8.11　分析结果 1

从图 8.10 的报表中，可以得出以下结论。

(1) 由修正的多重判定系数值为 0.524 082 636 可知，此 5 项因素与未来是否会购买啤酒之间存在回归关系，而回归模型的解释能力为中等，因为此值越接近 1，代表解释能力越强。

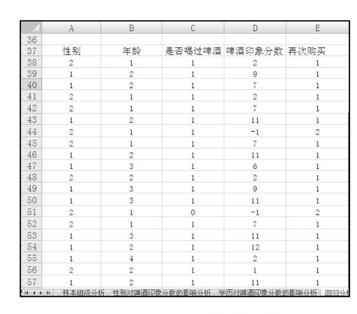
(2) 由方差分析检验结果来看，F 检验的统计量为 F=7.386 989 665，对应的 P 值为 0.000 256 788，这个值非常小，所以结论为拒绝原假设，接受备择假设，即表示 5 项因素与未来购买啤酒之间存在显著性的相关性。

(3) 各项系数的检验。各项系数的原假设和备择假设分别为 $H_0:\beta_i=0$；$H_1:\beta_i\neq 0$，检验的统计量为 t 检验，那么性别、年龄、是否喝过啤酒、最常喝的品牌、啤酒印象分数 5 项因素所对应的 P 值分别为 0.222 281 961、0.158 963 656、0.002 787 937、0.859 185 671、0.029 844 544。其中只有是否喝过啤酒和啤酒印象分数的 P 值小于 0.05，说明是否喝过啤酒和啤酒印象分数与未来是否购买啤酒存在相关性。这时选择消除 P 值最大的对应的变量，建立一个四元回归模型。

2. 建立一个四元线性回归预测模型

第一步：把“性别”、“年龄”、“是否喝过啤酒”、“啤酒印象分类”和“再次购买”数据复制到“回归分析“工作表的 A37:E67 单元格区域中，如图 8.12 所示。

第二步：单击“数据”→“分析”→“数据分析”按钮，弹出“数据分析”对话框，在“分析工具”列表中选择“回规”选项，如图 8.9 所示，然后单击“确定”按钮，弹出“回归”对话框，在“Y 值输入区域”文本框中输入“$E$37:$E$67”，在“X 值输入区域”文本框中输入“$A$37:$D$67”，选择“标志”和“置信度”复选框，其中的置信度默认为“95%”，选中“输出区域”单选按钮，并在其文本框中输入“$F$43”，如图 8.13 所示。



	A	B	C	D	E
36					
37	性别	年龄	是否喝过啤酒	啤酒印象分数	再次购买
38	2	1	1	2	1
39	1	2	1	9	1
40	1	2	1	7	1
41	2	1	1	2	1
42	2	1	1	7	1
43	1	2	1	11	1
44	2	1	1	-1	2
45	2	1	1	7	1
46	1	2	1	11	1
47	1	3	1	6	1
48	2	2	1	2	1
49	1	3	1	9	1
50	1	3	1	11	1
51	2	1	0	-1	2
52	2	1	1	7	1
53	1	3	1	11	1
54	1	2	1	12	1
55	1	4	1	2	1
56	2	2	1	1	1
57	1	2	1	11	1

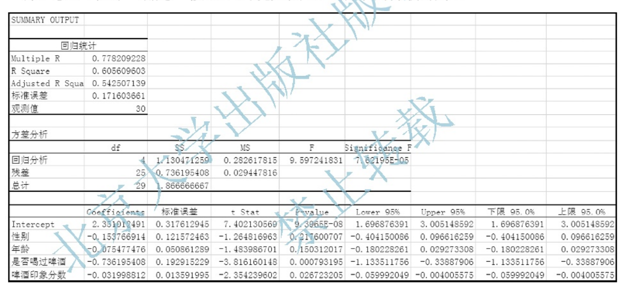
样本组成分析 / 性别对啤酒印象分数的影响分析 / 学历对啤酒印象分数的影响分析 / 回归分

图 8.12　A37:E67 单元格区域



图 8.13　“回归”对话框 2

第三步：单击“确定”按钮，出现如图 8.14 所示的分析结果。



SUMMARY OUTPUT

回归统计	
Multiple R	0.778209228
R Square	0.605609603
Adjusted R Squa	0.542507139
标准误差	0.171603661
观测值	30

方差分析

	df	SS	MS	F	Significance F
回归分析	4	1.130471259	0.282617815	9.597241831	7.62195E-05
残差	25	0.736195408	0.029447816		
总计	29	1.866666667			

	Coefficients	标准误差	t Stat	P-value	Lower 95%	Upper 95%	下限 95.0%	上限 95.0%
Intercept	2.351012491	0.317612945	7.402130569	9.3965E-08	1.696876391	3.005148592	1.696876391	3.005148592
性别	-0.153766914	0.121572463	-1.264816963	0.217600707	-0.404150086	0.096616259	-0.404150086	0.096616259
年龄	-0.075477476	0.050861289	-1.483986701	0.150312017	-0.180228261	0.029273308	-0.180228261	0.029273308
是否喝过啤酒	-0.736195408	0.192915229	-3.816160148	0.000793195	-1.133511756	-0.33887906	-1.133511756	-0.33887906
啤酒印象分数	-0.031998812	0.013591995	-2.354239602	0.026723205	-0.059992049	-0.004005575	-0.059992049	-0.004005575

图 8.14　分析结果 2

从图 8.14 的报表中，可以得出以下结论。

(1) 由修正的多重判定系数值为 0.542 507 139，可知此 4 项因素与未来是否会购买啤酒之间存在回归关系，而回归模型的解释能力为中等。

(2) 由方差分析检验结果来看，F 检验的统计量为 9.597 241 831，对应的 P 值为 7.62195×10^{-5}，这个值非常小，所以结论为拒绝原假设，接受备择假设，即表示 4 项因素与未来是否购买啤酒之间存在显著性的相关性。

(3) 各项系数的检验。检验的统计量为 t 检验，那么性别、年龄、是否喝过啤酒、啤酒印象分数 4 项因素所对应的 P 值分别为 0.217 600 707、0.150 312 017、0.000 793 195、0.026 723 205。同样只有是否喝过啤酒和啤酒印象分数的 P 值小于 0.05，说明是否喝过啤酒和啤酒印象分数与未来是否购买啤酒存在相关性。这时选择消除 P 值最大的对应的变量，建立一个三元回归模型。

3. 建立一个三元线性回归预测模型

第一步：把“年龄”“是否喝过啤酒”“啤酒印象”和“再次购买”数据复制到“回归分析“工作表的 A76:D106 单元格区域中，如图 8.15 所示。

	A	B	C	D	E
76	年龄	是否喝过啤酒	啤酒印象分数	再次购买	
77	1	1	2	1	
78	2	1	9	1	
79	2	1	7	1	
80	1	1	2	1	
81	1	1	7	1	
82	2	1	11	1	
83	1	1	-1	2	
84	1	1	7	1	
85	2	1	11	1	
86	3	1	6	1	
87	2	1	2	1	
88	3	1	9	1	
89	3	1	11	1	
90	1	0	-1	2	
91	1	1	7	1	
92	3	1	11	1	
93	2	1	12	1	
94	4	1	2	1	
95	2	1	1	1	
96	2	1	11	1	
97	1	1	11	1	

样本组成分析 | 性别对啤酒印象分数的影响分析 | 学历对啤酒印象分数的影响分析 | 回归分

图 8.15 A76:D106 单元格区域

第二步：单击“数据”→“分析”→“数据分析”按钮，弹出“数据分析”对话框，在“分析工具”列表中选择“回归”选项，如图 8.9 所示，然后单击“确定”按钮，弹出“回归”对话框，在“Y 值输入区域”文本框中输入“$D$76:$D$106”，在“X 值输入区域”文本框中输入“$A$76:$C$106”，选择“标志”和“置信度”复选框，其中的置信度默认为“95%”，选中“输出区域”单选按钮，并在其文本框中输入“$F$86”，如图 8.16 所示。



图 8.16 “回归”对话框 3

第三步：单击“确定”按钮，出现如图 8.17 所示的分析结果。

SUMMARY OUTPUT								
回归统计								
Multiple R	0.761821746							
R Square	0.580372373							
Adjusted R Squ	0.5319538							
标准误差	0.173571633							
观测值	30							
方差分析								
	df	SS	MS	F	ignificance F			
回归分析	3	1.083361763	0.361120588	11.98656516	4.0884E-05			
残差	26	0.783304904	0.030127112					
总计	29	1.866666667						
	Coefficients	标准误差	t Stat	P-value	Lower 95%	Upper 95%	下限 95.0%	上限 95.0%
Intercept	2.017200812	0.178728797	11.28637827	1.61091E-11	1.649818511	2.384583112	1.649818511	2.384583112
年龄	-0.03587624	0.040541629	-0.88492358	0.384308777	-0.11921075	0.047458267	-0.119210753	0.047458267
是否喝过啤酒	-0.7833049	0.191456285	-4.09129898	0.000368322	-1.17684893	-0.38976088	-1.17684893	-0.389760878
啤酒印象分数	-0.01867543	0.008688064	-2.14955029	0.041070926	-0.036534	-0.00081686	-0.036534003	-0.000816859

图 8.17　分析结果 3

从图 8.17 的报表中，可以得出以下结论。

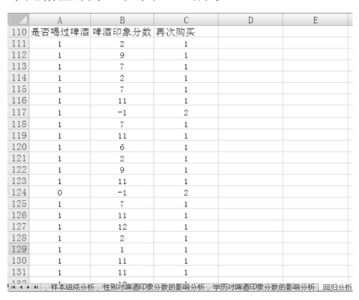
(1) 由修正的多重判定系数值为 0.531 953 8，可知此 3 项因素与未来是否会购买啤酒之间存在回归关系，而回归模型的解释能力为中等。

(2) 由方差分析检验结果来看，F 检验的统计量为 11.966 565 16，对应的 P 值为 $4.088\,4\times10^{-5}$，这个值非常小，所以结论为拒绝原假设，接受备择假设，即表示 3 项因素与未来是否购买啤酒之间存在显著性的相关性。

(3) 各项系数的检验。各项系数的原假设和备择假设分别为 $H_0:\beta_i=0$； $H_1:\beta_i\neq0$，检验的统计量为 t 检验，那么年龄、是否喝过啤酒、啤酒印象分数 3 项因素所对应的 P 值分别为 0.384 308 777、0.000 368 322、0.041 070 926。还是只有年龄的回归系数不显著，这时选择消除 P 值最大的对应的变量，建立一个二元回归模型。

4. 建立一个二元线性回归预测模型

第一步：把“是否喝过啤酒”“啤酒印象分数”和“再次购买”数据复制到“回归分析”工作表的 A110:C140 单元格区域中，如图 8.18 所示。



	A	B	C	D	E
110	是否喝过啤酒	啤酒印象分数	再次购买		
111	1	2	1		
112	1	9	1		
113	1	7	1		
114	1	2	1		
115	1	7	1		
116	1	11	1		
117	1	-1	2		
118	1	7	1		
119	1	11	1		
120	1	6	1		
121	1	2	1		
122	1	9	1		
123	1	11	1		
124	0	-1	2		
125	1	7	1		
126	1	11	1		
127	1	12	1		
128	1	2	1		
129	1	1	1		
130	1	11	1		
131	1	11	1		

样本组成分析 / 性别对啤酒印象分数的影响分析 / 学历对啤酒印象分数的影响分析 / 回归分析

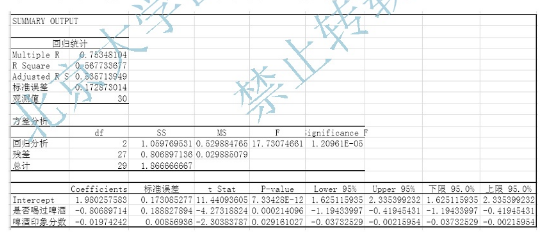
图 8.18　A110:C140 单元格区域

第二步：单击“数据”→“分析”→“数据分析”按钮，弹出“数据分析”对话框，在“分析工具”列表中选择“回归”选项，如图 8.9 所示，然后单击“确定”按钮，弹出“回归”对话框，在“Y 值输入区域”文本框中输入“\$C\$110:\$C\$140”，在“X 值输入区域”文本框中输入“\$A\$110:\$B\$140”，选中“标志”和“置信度”复选框，其中的置信度默认为“95%”，选中“输出区域”单选按钮，并在其文本框中输入“\$D\$118”，如图 8.19 所示。



图 8.19 “回归”对话框 4

第三步：单击“确定”按钮，出现如图 8.20 所示的分析结果。



SUMMARY OUTPUT

回归统计	
Multiple R	0.75348104
R Square	0.567733677
Adjusted R S	0.535713949
标准误差	0.172873014
观测值	30

方差分析

	df	SS	MS	F	Significance F
回归分析	2	1.059769531	0.529884765	17.73074661	1.20961E-05
残差	27	0.806897136	0.029885079		
总计	29	1.866666667			

	Coefficients	标准误差	t Stat	P-value	Lower 95%	Upper 95%	下限 95.0%	上限 95.0%
Intercept	1.980257583	0.173085277	11.44093605	7.33428E-12	1.625115935	2.335399232	1.625115935	2.335399232
是否喝过啤酒	-0.80689714	0.188827894	-4.27318824	0.000214096	-1.19433997	-0.41945431	-1.19433997	-0.41945431
啤酒印象分数	-0.01974242	0.00856936	-2.30383787	0.029161027	-0.03732529	-0.00215954	-0.03732529	-0.00215954

图 8.20 分析结果 4

从图 8.20 的报表中，可以得出以下结论。

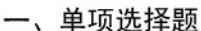
(1) 由修正的多重判定系数值为 0.535 713 949，可知此两项因素与未来是否会购买啤酒之间存在回归关系，而回归模型的解释能力为中等。

(2) 由方差分析检验结果来看，F 检验的统计量为 17.730 746 61，对应的 P 值为 $1.209\,61\times10^{-5}$，这个值非常小，所以结论为拒绝原假设，接受备择假设，即表示两项因素与未来购买啤酒之间存在显著性的相关性。

(3) 各项系数的检验。各项系数的原假设和备择假设分别为 $H_0:\beta_i=0$；$H_1:\beta_i\neq0$，检验的统计量为 t 检验，那么是否喝过啤酒、啤酒印象分数这两项因素所对应的 P 值分别

为 0.000 214 096，0.029 161 027，均小于 0.05，表示这两项因素与未来是否购买啤酒之间有非常显著性的关系。因此，该二元回归预测模型为最简单也是最具有解释能力的预测模型。

习　　题



一、单项选择题

1．变量之间存在的不确定的数量关系，称为(　　)。

A．相关关系　　B．函数关系　　C．线性关系　　D．非线性关系

2．两变量的样本之间的关系强度采用(　　)。

A．相关系数　　B．判定系数　　C．R^2　　D．回归系数

3．根据图 8.21，可以判断两个变量之间存在(　　)。

图 8.21　散点图

A．正线性相关关系　　B．负线性相关关系

C．非线性关系　　D．函数关系

4．根据相关系的特点，下面的相关系数取值错误的是(　　)。

A．0.89　　B．1.03　　C．−0.5　　D．0

5．下面关于相关系数叙述错误的是(　　)。

A．γ 的取值范围为：$-1 \leqslant \gamma \leqslant 1$

B．γ 具有对称性

C．γ 的大小与 x 和 y 的原点及尺度无关

D．γ 的取值范围为 $0 \leqslant \gamma \leqslant 1$

6．计算儿童身高和年龄的相关系数时，其中身高采用“cm”比“m”的相关系数(　　)。

A．增加　　B．减少　　C．不变　　D．不确定

7．Excel 中的相关系数函数为(　　)。

A．CORREL　　B．MODE　　C．STEDV　　D．AVERAGE

8．在相关关系的显著性检验中，原假设和备择假设为(　　)。

A．$H_0: \rho = 0, H_1: \rho \neq 0$　　B．$H_0: r = 0, H_1: r \neq 0$

C．$H_0: \rho \geqslant 0, H_1: \rho < 0$　　D．$H_0: \rho \leqslant 0, H_1: \rho > 0$

9．在相关系数的显著性检验中，检验的统计量是(　　)。

A．标准正态　　B．t　　C．F　　D．χ^2

10．样本的回归方程 $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ 中，参数估计值的估计方法是(　　)。

A．最小二乘法　　B．极大似然估计法

C．点估计　　D．矩估计

11．最小二乘法是指所有点到直线距离平方和最小，其中距离指的是(　　)。

A．竖直距离　　B．水平距离　　C．垂直距离　　D．以上都不对

12．样本的回归方程 $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ 中，$\hat{\beta}_0$ 和 $\hat{\beta}_1$ 的关系(　　)。

A．相关　　B．不相关　　C．不确定　　D．$\text{cov}(\hat{\beta}_0，\hat{\beta}_1) = 0$

13．判定系数 R^2 的取值范围为(　　)。

A．$[0,1]$　　B．$[-1,1]$　　C．$[-1,0]$　　D．$[0,+\infty]$

14．在一元线性回归分析中，总平方和服从 χ^2 分布，其自由度为(　　)。

A．$n-1$　　B．1　　C．2　　D．$n-2$

15．在线性回归分析中，回归平方和服从 χ^2 分布，其自由度为(　　)。

A．$n-1$　　B．1　　C．2　　D．$n-2$

16．在线性回归分析中，残差平方和服从 χ^2 分布，其自由度为(　　)。

A．$n-1$　　B．1　　C．2　　D．$n-2$

17．残差均方的平方根，称为(　　)。

A．估计标准误差　　B．回归均方

C．回归平方和　　D．残差平方和

18．估计标准误差是反映实际观测值 y_i 与样本回归方程的估计值 $\hat{y}_i$ 之间的差异的大小，其中 s_e 为(　　)时，代表样本回归方程拟合得最好。

A．0　　B．1　　C．$+\infty$　　D．$-\infty$

19．在线性关系检验中检验的统计量是(　　)分布。

A．标准正态　　B．t　　C．F　　D．χ^2

20．在一元线性回归分析中，回归系数检验的统计量是(　　)分布。

A．标准正态　　B．t　　C．F　　D．χ^2

二、多项选择题

1．对样本回归方程进行评价的指标有(　　)。

A．R^2　　B．估计标准误差　　C．SSR　　D．SST

2．在一元回归模型 $y = \beta_0 + \beta_1 x + \xi$ 中，其中误差项 ξ 的满足条件有(　　)。

A．$\xi \sim N(0,\sigma^2)$　　B．$\text{cov}(x,\xi) = 0$　　C．x 是非随机的　　D．$\text{cov}(x,\xi) \neq 0$

3．判定系数 R^2 的计算公式为(　　)。

A．$R^2 = \dfrac{\text{SSR}}{\text{SST}}$　　B．$R^2 = 1 - \dfrac{\text{SSE}}{\text{SST}}$　　C．$R^2 = \dfrac{\text{SSE}}{\text{SST}}$　　D．$R^2 = \dfrac{\text{SSR}}{\text{SSE}}$

4．回归方程的统计检验主要有(　　)。

A．线性关系的检验　　B．回归系数的检验

C．判定系数的检验　　D．相关系数的检验

5．根据估计方程进行估计和预测的方法，主要包括(　　)。

A．点估计　　B．区间估计

C．平均值的点估计　　D．个别值的点估计

三、简答题

1．简述经济预测和经济控制的步骤。

2．简述相关分析的步骤。

3．以一元线性回归为例，简述最小二乘法的思路。

4．以一元线性回归为例，简述线性关系检验和回归系数检验的步骤。

四、计算题

1．一家超市集团拥有多家子超市，公司的管理者想通过广告支出来估计销售收入，为此他随机抽取了 7 家子超市，得到广告支出和销售收入的数据见表 8-2 所示。

表 8-2　广告支出和销售收入数据

单位：万元

广告费用支出	销售额
1	19
2	32
4	44
6	40
10	52
14	53
20	54

问题：

(1) 试画出广告费用支出与销售额的散点图。

(2) 计算广告费用支出与销售额的关系强度。

(3) 对广告费用支出与销售额进行相关系数检验。

(4) 利用最小二乘法，计算样广告费用支出与销售额的样本回归方程。

(5) 评价样本回归方程。

(6) 写出广告费用支出与销售额进行线性关系和回归系数检验过程。

2．根据某 8 个地区的人均可支配收入(y)与人均消费水平(x)的数据，得到 8-3 和表 8-4 所示的结果。($\alpha = 0.05$)

表 8-3　方差分析表

变差来源	自由度	平方和	均方	*F*	*P*	*F*-crit
回归					8.481 19E-06	5.987 4
残差		53 845.02		—	—	—
总计		1 798 550.42	—	—	—	—

表 8-4　参数估计表

	Cofficients	标准误差	*t* Stat	*P*
Intercep	5 050.5	867.783 5		0.000
X Variable	0.658		19.98	2.17E-09

(1) 完成方差分析表和参数估计表。

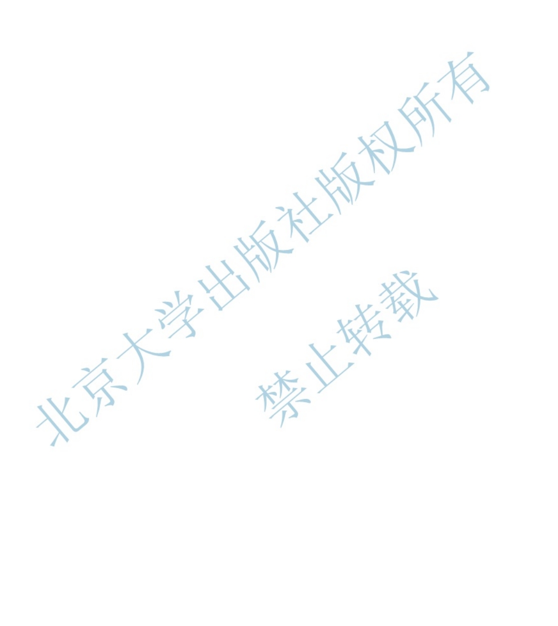
(2) 根据参数估计表，写出回归方程。

(3) 根据方差分析表，评价样本回归方程。

(4) 写出线性关系检验的步骤。

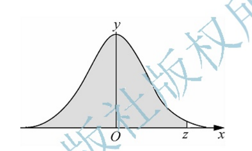
(5) 写出自变量回归系数的检验步骤。

(6) 其中某个地区的人均可支配收入为 27 873 元，预测该地区的人均消费水平的置信区间估计和预测区间估计。



附录　用 Excel 生成概率分布表

附表 1　标准正态分布表



利用 Excel 提供的统计函数“NORMSDIST”可以生成标准正态分布的累积概率分布表，即 $P(Z \leqslant x)$。生成标准正态分布累积概率分布表可按以下步骤进行。

第一步：将 x 的值（可由读者需要自行确定）输入到工作表的 A 列，将 x 取值的尾数输入到第 1 行，形成标准正态分布的表头，如下图所示：

	A	B	C	D	E	F	G	H	I	J	K
1	x	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
2	0.0										
3	0.1										
4	0.2										
5	0.3										
6	0.4										
7	0.5										
8	0.6										
9	0.7										
10	0.8										

第二步：在 B2 单元格输入公式“=NORMSDIST($A2+B$1)”，其余结果可通过向下、向右复制而得到。可根据需要生成不同 x 的标准正态分布概率表，现将按照上述方法生成的部分结果展示如下表：

	A	B	C	D	E	F	G	H	I	J	K
1	x	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
2	0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
3	0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
4	0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
5	0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
6	0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
7	0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
8	0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
9	0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
10	0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
11	0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
12	1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
13	1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
14	1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
15	1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
16	1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
17	1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
18	1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
19	1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
20	1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
21	1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
22	2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
23	2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
24	2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
25	2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
26	2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
27	2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
28	2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
29	2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
30	2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
31	2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
32	3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990
33	3.1	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
34	3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995
35	3.3	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997
36	3.4	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998
37	3.5	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998
38	3.6	0.9998	0.9998	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999
39	3.7	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999
40	3.8	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999
41	3.9	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
42	4.0	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
43	4.1	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
44	4.2	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
45	4.3	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
46	4.4	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
47	4.5	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
48	4.6	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
49	4.7	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
50	4.8	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
51	4.9	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

附表 2　标准正态分布临界值表

利用 Excel 提供的统计函数“NORMSINV”，可以生成标准正态分布的临界值表，临界值是根据标准正态分布随机变量分布的累积概率的值计算的。如果有 $P(Z \leqslant x) = p$，则对于任意给定的 $p(0 \leqslant p \leqslant 1)$ 可以求出相应的 x。用 Excel 生成标准正态分布临界值表可进行如下操作。

第一步：将标准正态变量累积概率的值输入到工作表的 A 列，其尾数输入到第一行，形成标准正态分布临界值表的表头，如下图所示：

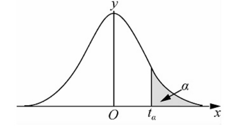
	A	B	C	D	E	F	G	H	I	J	K
1	x	0.000	0.001	0.002	0.003	0.004	0.005	0.006	0.007	0.008	0.009
2	0.50										
3	0.51										
4	0.52										
5	0.53										
6	0.54										
7	0.55										
8	0.56										
9	0.57										
10	0.58										
11	0.59										
12	0.60										
13	0.61										
14	0.62										
15	0.63										
16	0.64										
17	0.65										

第二步：在 B2 单元格输入公式“=NORMSINV($A2+B$1)”，其它结果通过向下、向右复制即可得到。可根据需要生成不同 p 值的标准正态分布临界值表，按照上述步骤操作得到该表部分结果如下图所示：

	A	B	C	D	E	F	G	H	I	J	K
1	x	0.000	0.001	0.002	0.003	0.004	0.005	0.006	0.007	0.008	0.009
2	0.50	0.0000	0.0025	0.0050	0.0075	0.0100	0.0125	0.0150	0.0175	0.0201	0.0226
3	0.51	0.0251	0.0276	0.0301	0.0326	0.0351	0.0376	0.0401	0.0426	0.0451	0.0476
4	0.52	0.0502	0.0527	0.0552	0.0577	0.0602	0.0627	0.0652	0.0677	0.0702	0.0728
5	0.53	0.0753	0.0778	0.0803	0.0828	0.0853	0.0878	0.0904	0.0929	0.0954	0.0979
6	0.54	0.1004	0.1030	0.1055	0.1080	0.1105	0.1130	0.1156	0.1181	0.1206	0.1231
7	0.55	0.1257	0.1282	0.1307	0.1332	0.1358	0.1383	0.1408	0.1434	0.1459	0.1484
8	0.56	0.1510	0.1535	0.1560	0.1586	0.1611	0.1637	0.1662	0.1687	0.1713	0.1738
9	0.57	0.1764	0.1789	0.1815	0.1840	0.1866	0.1891	0.1917	0.1942	0.1968	0.1993
10	0.58	0.2019	0.2045	0.2070	0.2096	0.2121	0.2147	0.2173	0.2198	0.2224	0.2250
11	0.59	0.2275	0.2301	0.2327	0.2353	0.2378	0.2404	0.2430	0.2456	0.2482	0.2508
12	0.60	0.2533	0.2559	0.2585	0.2611	0.2637	0.2663	0.2689	0.2715	0.2741	0.2767
13	0.61	0.2793	0.2819	0.2845	0.2871	0.2898	0.2924	0.2950	0.2976	0.3002	0.3029

14	0.62	0.3055	0.3081	0.3107	0.3134	0.3160	0.3186	0.3213	0.3239	0.3266	0.3292
15	0.63	0.3319	0.3345	0.3372	0.3398	0.3425	0.3451	0.3478	0.3505	0.3531	0.3558
16	0.64	0.3585	0.3611	0.3638	0.3665	0.3692	0.3719	0.3745	0.3772	0.3799	0.3826
17	0.65	0.3853	0.3880	0.3907	0.3934	0.3961	0.3989	0.4016	0.4043	0.4070	0.4097
18	0.66	0.4125	0.4152	0.4179	0.4207	0.4234	0.4261	0.4289	0.4316	0.4344	0.4372
19	0.67	0.4399	0.4427	0.4454	0.4482	0.4510	0.4538	0.4565	0.4593	0.4621	0.4649
20	0.68	0.4677	0.4705	0.4733	0.4761	0.4789	0.4817	0.4845	0.4874	0.4902	0.4930
21	0.69	0.4959	0.4987	0.5015	0.5044	0.5072	0.5101	0.5129	0.5158	0.5187	0.5215
22	0.70	0.5244	0.5273	0.5302	0.5330	0.5359	0.5388	0.5417	0.5446	0.5476	0.5505
23	0.71	0.5534	0.5563	0.5592	0.5622	0.5651	0.5681	0.5710	0.5740	0.5769	0.5799
24	0.72	0.5828	0.5858	0.5888	0.5918	0.5948	0.5978	0.6008	0.6038	0.6068	0.6098
25	0.73	0.6128	0.6158	0.6189	0.6219	0.6250	0.6280	0.6311	0.6341	0.6372	0.6403
26	0.74	0.6433	0.6464	0.6495	0.6526	0.6557	0.6588	0.6620	0.6651	0.6682	0.6713
27	0.75	0.6745	0.6776	0.6808	0.6840	0.6871	0.6903	0.6935	0.6967	0.6999	0.7031
28	0.76	0.7063	0.7095	0.7128	0.7160	0.7192	0.7225	0.7257	0.7290	0.7323	0.7356
29	0.77	0.7388	0.7421	0.7454	0.7488	0.7521	0.7554	0.7588	0.7621	0.7655	0.7688
30	0.78	0.7722	0.7756	0.7790	0.7824	0.7858	0.7892	0.7926	0.7961	0.7995	0.8030
31	0.79	0.8064	0.8099	0.8134	0.8169	0.8204	0.8239	0.8274	0.8310	0.8345	0.8381
32	0.80	0.8416	0.8452	0.8488	0.8524	0.8560	0.8596	0.8633	0.8669	0.8705	0.8742
33	0.81	0.8779	0.8816	0.8853	0.8890	0.8927	0.8965	0.9002	0.9040	0.9078	0.9116
34	0.82	0.9154	0.9192	0.9230	0.9269	0.9307	0.9346	0.9385	0.9424	0.9463	0.9502
35	0.83	0.9542	0.9581	0.9621	0.9661	0.9701	0.9741	0.9782	0.9822	0.9863	0.9904
36	0.84	0.9945	0.9986	1.0027	1.0069	1.0110	1.0152	1.0194	1.0237	1.0279	1.0322
37	0.85	1.0364	1.0407	1.0450	1.0494	1.0537	1.0581	1.0625	1.0669	1.0714	1.0758
38	0.86	1.0803	1.0848	1.0893	1.0939	1.0985	1.1031	1.1077	1.1123	1.1170	1.1217
39	0.87	1.1264	1.1311	1.1359	1.1407	1.1455	1.1503	1.1552	1.1601	1.1650	1.1700
40	0.88	1.1750	1.1800	1.1850	1.1901	1.1952	1.2004	1.2055	1.2107	1.2160	1.2212
41	0.89	1.2265	1.2319	1.2372	1.2426	1.2481	1.2536	1.2591	1.2646	1.2702	1.2759
42	0.90	1.2816	1.2873	1.2930	1.2988	1.3047	1.3106	1.3165	1.3225	1.3285	1.3346
43	0.91	1.3408	1.3469	1.3532	1.3595	1.3658	1.3722	1.3787	1.3852	1.3917	1.3984
44	0.92	1.4051	1.4118	1.4187	1.4255	1.4325	1.4395	1.4466	1.4538	1.4611	1.4684
45	0.93	1.4758	1.4833	1.4909	1.4985	1.5063	1.5141	1.5220	1.5301	1.5382	1.5464
46	0.94	1.5548	1.5632	1.5718	1.5805	1.5893	1.5982	1.6072	1.6164	1.6258	1.6352
47	0.95	1.6449	1.6546	1.6646	1.6747	1.6849	1.6954	1.7060	1.7169	1.7279	1.7392
48	0.96	1.7507	1.7624	1.7744	1.7866	1.7991	1.8119	1.8250	1.8384	1.8522	1.8663
49	0.97	1.8808	1.8957	1.9110	1.9268	1.9431	1.9600	1.9774	1.9954	2.0141	2.0335
50	0.98	2.0537	2.0749	2.0969	2.1201	2.1444	2.1701	2.1973	2.2262	2.2571	2.2904
51	0.99	2.3263	2.3656	2.4089	2.4573	2.5121	2.5758	2.6521	2.7478	2.8782	3.0902

附表 3　t 分布临界值表



利用 Excel 提供的统计函数“TINV”可以生成 t 分布的临界值表，该表是根据 t 分布的右尾概率 α 计算的相应的临界值。如果 $P(t \geqslant x) = \alpha$，则对于任意给定的概率 $p(0 \leqslant \alpha \geqslant 1)$，

可以求出相应的 x。生成 t 分布临界值表的具体操作步骤如下。

第一步：在工作表 A 列中输入 t 分布自由度 df 的值，在第 1 行中输入右尾概率 α 的取值，构建出 t 分布临界值表的表头，如下图所示：

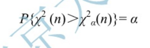
	A	B	C	D	E	F	G	H	I	J	K
1	df/α	0.250	0.200	0.150	0.100	0.050	0.025	0.010	0.005	0.001	0.0005
2	1										
3	2										
4	3										
5	4										
6	5										
7	6										
8	7										
9	8										
10	9										
11	10										
12	11										
13	12										
14	13										
15	14										
16	15										

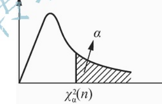
第二步：在 B2 单元格输入公式“=TINV(B$1*2，$A2)”，并将其向下、向右复制即可得到 t 分布的临界值表，读者可根据需要生成不同 α 和不同自由度的 t 分布的临界值表，现将按照上述步骤操作得到的部分结果展示如下：

	A	B	C	D	E	F	G	H	I	J	K
1	df/α	0.250	0.200	0.150	0.100	0.050	0.025	0.010	0.005	0.001	0.0005
2	1	1.0000	1.3764	1.9626	3.0777	6.3138	12.7062	31.8205	63.6567	318.3088	636.6192
3	2	0.8165	1.0607	1.3862	1.8856	2.9200	4.3027	6.9646	9.9248	22.3271	31.5991
4	3	0.7649	0.9785	1.2498	1.6377	2.3534	3.1824	4.5407	5.8409	10.2145	12.9240
5	4	0.7407	0.9410	1.1896	1.5332	2.1318	2.7764	3.7469	4.6041	7.1732	8.6103
6	5	0.7267	0.9195	1.1558	1.4759	2.0150	2.5706	3.3649	4.0321	5.8934	6.8688
7	6	0.7176	0.9057	1.1342	1.4398	1.9432	2.4469	3.1427	3.7074	5.2076	5.9588
8	7	0.7111	0.8960	1.1192	1.4149	1.8946	2.3646	2.9980	3.4995	4.7853	5.4079
9	8	0.7064	0.8889	1.1081	1.3968	1.8595	2.3060	2.8965	3.3554	4.5008	5.0413
10	9	0.7027	0.8834	1.0997	1.3830	1.8331	2.2622	2.8214	3.2498	4.2968	4.7809
11	10	0.6998	0.8791	1.0931	1.3722	1.8125	2.2281	2.7638	3.1693	4.1437	4.5869
12	11	0.6974	0.8755	1.0877	1.3634	1.7959	2.2010	2.7181	3.1058	4.0247	4.4370
13	12	0.6955	0.8726	1.0832	1.3562	1.7823	2.1788	2.6810	3.0545	3.9296	4.3178
14	13	0.6938	0.8702	1.0795	1.3502	1.7709	2.1604	2.6503	3.0123	3.8520	4.2208
15	14	0.6924	0.8681	1.0763	1.3450	1.7613	2.1448	2.6245	2.9768	3.7874	4.1405
16	15	0.6912	0.8662	1.0735	1.3406	1.7531	2.1314	2.6025	2.9467	3.7328	4.0728
17	16	0.6901	0.8647	1.0711	1.3368	1.7459	2.1199	2.5835	2.9208	3.6862	4.0150
18	17	0.6892	0.8633	1.0690	1.3334	1.7396	2.1098	2.5669	2.8982	3.6458	3.9651
19	18	0.6884	0.8620	1.0672	1.3304	1.7341	2.1009	2.5524	2.8784	3.6105	3.9216
20	19	0.6876	0.8610	1.0655	1.3277	1.7291	2.0930	2.5395	2.8609	3.5794	3.8834
21	20	0.6870	0.8600	1.0640	1.3253	1.7247	2.0860	2.5280	2.8453	3.5518	3.8495
22	21	0.6864	0.8591	1.0627	1.3232	1.7207	2.0796	2.5176	2.8314	3.5272	3.8193
23	22	0.6858	0.8583	1.0614	1.3212	1.7171	2.0739	2.5083	2.8188	3.5050	3.7921
24	23	0.6853	0.8575	1.0603	1.3195	1.7139	2.0687	2.4999	2.8073	3.4850	3.7676

25	24	0.6848	0.8569	1.0593	1.3178	1.7109	2.0639	2.4922	2.7969	3.4668	3.7454
26	25	0.6844	0.8562	1.0584	1.3163	1.7081	2.0595	2.4851	2.7874	3.4502	3.7251
27	26	0.6840	0.8557	1.0575	1.3150	1.7056	2.0555	2.4786	2.7787	3.4350	3.7066
28	27	0.6837	0.8551	1.0567	1.3137	1.7033	2.0518	2.4727	2.7707	3.4210	3.6896
29	28	0.6834	0.8546	1.0560	1.3125	1.7011	2.0484	2.4671	2.7633	3.4082	3.6739
30	29	0.6830	0.8542	1.0553	1.3114	1.6991	2.0452	2.4620	2.7564	3.3962	3.6594
31	30	0.6828	0.8538	1.0547	1.3104	1.6973	2.0423	2.4573	2.7500	3.3852	3.6460
32	31	0.6825	0.8534	1.0541	1.3095	1.6955	2.0395	2.4528	2.7440	3.3749	3.6335
33	32	0.6822	0.8530	1.0535	1.3086	1.6939	2.0369	2.4487	2.7385	3.3653	3.6218
34	33	0.6820	0.8526	1.0530	1.3077	1.6924	2.0345	2.4448	2.7333	3.3563	3.6109
35	34	0.6818	0.8523	1.0525	1.3070	1.6909	2.0322	2.4411	2.7284	3.3479	3.6007
36	35	0.6816	0.8520	1.0520	1.3062	1.6896	2.0301	2.4377	2.7238	3.3400	3.5911
37	36	0.6814	0.8517	1.0516	1.3055	1.6883	2.0281	2.4345	2.7195	3.3326	3.5821
38	37	0.6812	0.8514	1.0512	1.3049	1.6871	2.0262	2.4314	2.7154	3.3256	3.5737
39	38	0.6810	0.8512	1.0508	1.3042	1.6860	2.0244	2.4286	2.7116	3.3190	3.5657
40	39	0.6808	0.8509	1.0504	1.3036	1.6849	2.0227	2.4258	2.7079	3.3128	3.5581
41	40	0.6807	0.8507	1.0500	1.3031	1.6839	2.0211	2.4233	2.7045	3.3069	3.5510
42	41	0.6805	0.8505	1.0497	1.3025	1.6829	2.0195	2.4208	2.7012	3.3013	3.5442
43	42	0.6804	0.8503	1.0494	1.3020	1.6820	2.0181	2.4185	2.6981	3.2960	3.5377
44	43	0.6802	0.8501	1.0491	1.3016	1.6811	2.0167	2.4163	2.6951	3.2909	3.5316
45	44	0.6801	0.8499	1.0488	1.3011	1.6802	2.0154	2.4141	2.6923	3.2861	3.5258
46	45	0.6800	0.8497	1.0485	1.3006	1.6794	2.0141	2.4121	2.6896	3.2815	3.5203
47	46	0.6799	0.8495	1.0483	1.3002	1.6787	2.0129	2.4102	2.6870	3.2771	3.5150
48	47	0.6797	0.8493	1.0480	1.2998	1.6779	2.0117	2.4083	2.6846	3.2729	3.5099
49	48	0.6796	0.8492	1.0478	1.2994	1.6772	2.0106	2.4066	2.6822	3.2689	3.5051
50	49	0.6795	0.8490	1.0475	1.2991	1.6766	2.0096	2.4049	2.6800	3.2651	3.5004

附表 4 χ^2 分布表





α	0.995	0.99	0.975	0.95	0.90	0.75
1	0.000	0.000	0.001	0.004	0.016	0.102
2	0.010	0.020	0.051	0.103	0.211	0.575
3	0.072	0.115	0.216	0.352	0.584	1.213
4	0.207	0.297	0.484	0.711	1.064	1.923
5	0.412	0.554	0.831	1.145	1.610	2.675
6	0.676	0.872	1.237	1.635	2.204	3.455
7	0.989	1.239	1.690	2.167	2.833	4.255
8	1.344	1.647	2.180	2.733	3.490	5.071
9	1.735	2.088	2.700	3.325	4.168	5.899
10	2.156	2.558	3.247	3.940	4.865	6.737
11	2.603	3.053	3.816	4.575	5.578	7.584
12	3.074	3.571	4.404	5.226	6.304	8.438
13	3.565	4.107	5.009	5.892	7.041	9.299
14	4.075	4.660	5.629	6.571	7.790	10.165
15	4.601	5.229	6.262	7.261	8.547	11.037

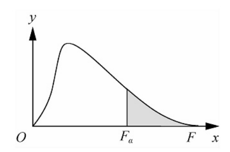
续表

α	0.995	0.99	0.975	0.95	0.90	0.75
16	5.142	5.812	6.908	7.962	9.312	11.912
17	5.697	6.408	7.564	8.672	10.085	12.792
18	6.265	7.015	8.231	9.390	10.865	13.675
19	6.844	7.633	8.907	10.117	11.651	14.562
20	7.434	8.260	9.591	10.851	12.443	15.452
21	8.034	8.897	10.283	11.591	13.240	16.344
22	8.643	9.542	10.982	12.338	14.041	17.240
23	9.260	10.196	11.689	13.091	14.848	18.137
24	9.886	10.856	12.401	13.848	15.659	19.037
25	10.520	11.524	13.120	14.611	16.473	19.939
26	11.160	12.198	13.844	15.379	17.292	20.843
27	11.808	12.878	14.573	16.151	18.114	21.749
28	12.461	13.565	15.308	16.928	18.939	22.657
29	13.121	14.256	16.047	17.708	19.768	23.567
30	13.787	14.953	16.791	18.493	20.599	24.478
31	14.458	15.655	17.539	19.281	21.434	25.390
32	15.134	16.362	18.291	20.072	22.271	26.304
33	15.815	17.073	19.047	20.867	23.110	27.219
34	16.501	17.789	19.806	21.664	23.952	28.136
35	17.192	18.509	20.569	22.465	24.797	29.054
36	17.887	19.233	21.336	23.269	25.643	29.973
37	18.586	19.960	22.106	24.075	26.492	30.893
38	19.289	20.691	22.878	24.884	27.343	31.815
39	19.996	21.426	23.654	25.695	28.196	32.737
40	20.707	22.164	24.433	26.509	29.051	33.660
41	21.421	22.906	25.215	27.326	29.907	34.585
42	22.138	23.650	25.999	28.144	30.765	35.510
43	22.860	24.398	26.785	28.965	31.625	36.436
44	23.584	25.148	27.575	29.787	32.487	37.363
45	24.311	25.901	28.366	30.612	33.350	38.291
α	0.25	0.1	0.05	0.025	0.01	0.005
1	1.323	2.706	3.841	5.024	6.635	7.879
2	2.773	4.605	5.991	7.378	9.210	10.597
3	4.108	6.251	7.815	9.348	11.345	12.838
4	5.385	7.779	9.488	11.143	13.277	14.860
5	6.626	9.236	11.070	12.832	15.086	16.750
6	7.841	10.645	12.592	14.449	16.812	18.548
7	9.037	12.017	14.067	16.013	18.475	20.278

续表

α	0.25	0.1	0.05	0.025	0.01	0.005
8	10.219	13.362	15.507	17.535	20.090	21.955
9	11.389	14.684	16.919	19.023	21.666	23.589
10	12.549	15.987	18.307	20.483	23.209	25.188
11	13.701	17.275	19.675	21.920	24.725	26.757
12	14.845	18.549	21.026	23.337	26.217	28.300
13	15.984	19.812	22.362	24.736	27.688	29.819
14	17.117	21.064	23.685	26.119	29.141	31.319
15	18.245	22.307	24.996	27.488	30.578	32.801
16	19.369	23.542	26.296	28.845	32.000	34.267
17	20.489	24.769	27.587	30.191	33.409	35.718
18	21.605	25.989	28.869	31.526	34.805	37.156
19	22.718	27.204	30.144	32.852	36.191	38.582
20	23.828	28.412	31.410	34.170	37.566	39.997
21	24.935	29.615	32.671	35.479	38.932	41.401
22	26.039	30.813	33.924	36.781	40.289	42.796
23	27.141	32.007	35.172	38.076	41.638	44.181
24	28.241	33.196	36.415	39.364	42.980	45.558
25	29.339	34.382	37.652	40.646	44.314	46.928
26	30.435	35.563	38.885	41.923	45.642	48.290
27	31.528	36.741	40.113	43.195	46.963	49.645
28	32.620	37.916	41.337	44.461	48.278	50.994
29	33.711	39.087	42.557	45.722	49.588	52.335
30	34.800	40.256	43.773	46.979	50.892	53.672
31	35.887	41.422	44.985	48.232	52.191	55.002
32	36.973	42.585	46.194	49.480	53.486	56.328
33	38.058	43.745	47.400	50.725	54.775	57.648
34	39.141	44.903	48.602	51.966	56.061	58.964
35	40.223	46.059	49.802	53.203	57.342	60.275
36	41.304	47.212	50.998	54.437	58.619	61.581
37	42.383	48.363	52.192	55.668	59.893	62.883
38	43.462	49.513	53.384	56.895	61.162	64.181
39	44.539	50.660	54.572	58.120	62.428	65.475
40	45.616	51.805	55.758	59.342	63.691	66.766
41	46.692	52.949	56.942	60.561	64.950	68.053
42	47.766	54.090	58.124	61.777	66.206	69.336
43	48.840	55.230	59.304	62.990	67.459	70.616
44	49.913	56.369	60.481	64.201	68.710	71.892
45	50.985	57.505	61.656	65.410	69.957	73.166

附表 5　F 分布临界值表



利用 Excel 提供的统计函数“FINV”可以生成 F 分布的临界值表，该表是根据 F 分布的右尾概率 α 计算的相应的临界值，即如果 $P(F \geqslant x) = \alpha$，则对于任意给定的概率 $p(0 \leqslant \alpha \leqslant 1)$，可以求出相应的 x。可按照如下操作步骤生成 F 分布临界值表。

第一步：将 F 分布右尾概率 α 的取值（如 $\alpha = 0.05$）输入到 B1 单元格中，将分子自由度 df1 的值输入到第 2 行中，将分母自由度 df2 的值输入在第 1 列中，如下图所示：

	A	B	C	D	E	F	G	H	I	J	K
1	α=	0.05									
2	*df2/df1*	1	2	3	4	5	6	7	8	9	10
3	1										
4	2										
5	3										
6	4										
7	5										
8	6										
9	7										
10	8										
11	9										
12	10										

第二步：在 B3 单元格输入公式“=FINV($B$1,B$2,$A3)”，并将其向下、向右复制即可得到 F 分布的临界值表，可根据需要生成不同 α 和不同自由度的 F 分布的临界值表，现以 $\alpha = 0.05$ 为例将 F 分布临界值表的部分结果展示如下：

	A	B	C	D	E	F	G	H	I	J	K
1	α=	0.05									
2	df2/df1	1	2	3	4	5	6	7	8	9	10
3	1	161.448	199.500	215.707	224.583	230.162	233.986	236.768	238.883	240.543	241.882
4	2	18.513	19.000	19.164	19.247	19.296	19.330	19.353	19.371	19.385	19.396
5	3	10.128	9.552	9.277	9.117	9.013	8.941	8.887	8.845	8.812	8.786
6	4	7.709	6.944	6.591	6.388	6.256	6.163	6.094	6.041	5.999	5.964
7	5	6.608	5.786	5.409	5.192	5.050	4.950	4.876	4.818	4.772	4.735
8	6	5.987	5.143	4.757	4.534	4.387	4.284	4.207	4.147	4.099	4.060
9	7	5.591	4.737	4.347	4.120	3.972	3.866	3.787	3.726	3.677	3.637
10	8	5.318	4.459	4.066	3.838	3.687	3.581	3.500	3.438	3.388	3.347
11	9	5.117	4.256	3.863	3.633	3.482	3.374	3.293	3.230	3.179	3.137
12	10	4.965	4.103	3.708	3.478	3.326	3.217	3.135	3.072	3.020	2.978
13	11	4.844	3.982	3.587	3.357	3.204	3.095	3.012	2.948	2.896	2.854
14	12	4.747	3.885	3.490	3.259	3.106	2.996	2.913	2.849	2.796	2.753
15	13	4.667	3.806	3.411	3.179	3.025	2.915	2.832	2.767	2.714	2.671
16	14	4.600	3.739	3.344	3.112	2.958	2.848	2.764	2.699	2.646	2.602
17	15	4.543	3.682	3.287	3.056	2.901	2.790	2.707	2.641	2.588	2.544
18	16	4.494	3.634	3.239	3.007	2.852	2.741	2.657	2.591	2.538	2.494
19	17	4.451	3.592	3.197	2.965	2.810	2.699	2.614	2.548	2.494	2.450
20	18	4.414	3.555	3.160	2.928	2.773	2.661	2.577	2.510	2.456	2.412
21	19	4.381	3.522	3.127	2.895	2.740	2.628	2.544	2.477	2.423	2.378
22	20	4.351	3.493	3.098	2.866	2.711	2.599	2.514	2.447	2.393	2.348
23	21	4.325	3.467	3.072	2.840	2.685	2.573	2.488	2.420	2.366	2.321
24	22	4.301	3.443	3.049	2.817	2.661	2.549	2.464	2.397	2.342	2.297
25	23	4.279	3.422	3.028	2.796	2.640	2.528	2.442	2.375	2.320	2.275
26	24	4.260	3.403	3.009	2.776	2.621	2.508	2.423	2.355	2.300	2.255
27	25	4.242	3.385	2.991	2.759	2.603	2.490	2.405	2.337	2.282	2.236
28	26	4.225	3.369	2.975	2.743	2.587	2.474	2.388	2.321	2.265	2.220
29	27	4.210	3.354	2.960	2.728	2.572	2.459	2.373	2.305	2.250	2.204
30	28	4.196	3.340	2.947	2.714	2.558	2.445	2.359	2.291	2.236	2.190
31	29	4.183	3.328	2.934	2.701	2.545	2.432	2.346	2.278	2.223	2.177
32	30	4.171	3.316	2.922	2.690	2.534	2.421	2.334	2.266	2.211	2.165
33	31	4.160	3.305	2.911	2.679	2.523	2.409	2.323	2.255	2.199	2.153
34	32	4.149	3.295	2.901	2.668	2.512	2.399	2.313	2.244	2.189	2.142
35	33	4.139	3.285	2.892	2.659	2.503	2.389	2.303	2.235	2.179	2.133
36	34	4.130	3.276	2.883	2.650	2.494	2.380	2.294	2.225	2.170	2.123
37	35	4.121	3.267	2.874	2.641	2.485	2.372	2.285	2.217	2.161	2.114
38	36	4.113	3.259	2.866	2.634	2.477	2.364	2.277	2.209	2.153	2.106
39	37	4.105	3.252	2.859	2.626	2.470	2.356	2.270	2.201	2.145	2.098
40	38	4.098	3.245	2.852	2.619	2.463	2.349	2.262	2.194	2.138	2.091
41	39	4.091	3.238	2.845	2.612	2.456	2.342	2.255	2.187	2.131	2.084
42	40	4.085	3.232	2.839	2.606	2.449	2.336	2.249	2.180	2.124	2.077
43	41	4.079	3.226	2.833	2.600	2.443	2.330	2.243	2.174	2.118	2.071
44	42	4.073	3.220	2.827	2.594	2.438	2.324	2.237	2.168	2.112	2.065
45	43	4.067	3.214	2.822	2.589	2.432	2.318	2.232	2.163	2.106	2.059
46	44	4.062	3.209	2.816	2.584	2.427	2.313	2.226	2.157	2.101	2.054
47	45	4.057	3.204	2.812	2.579	2.422	2.308	2.221	2.152	2.096	2.049
48	46	4.052	3.200	2.807	2.574	2.417	2.304	2.216	2.147	2.091	2.044
49	47	4.047	3.195	2.802	2.570	2.413	2.299	2.212	2.143	2.086	2.039
50	48	4.043	3.191	2.798	2.565	2.409	2.295	2.207	2.138	2.082	2.035
51	49	4.038	3.187	2.794	2.561	2.404	2.290	2.203	2.134	2.077	2.030
52	50	4.034	3.183	2.790	2.557	2.400	2.286	2.199	2.130	2.073	2.026

习 题 答 案

第 1 章　总论

1．

(1) 数值数据　(2) 数值数据　(3) 数值数据　(4) 分类数据　(5) 分类数据

(6) 分类数据　(7) 顺序数据

2．

(1) 总体是 IT 从业者；样本是 200IT 从业者；样本容量为 200。

(2) 数值数据；分类数据。

3．推断统计：(1)和(3)；描述性统计：(2)。

4．参数：32%；统计量 35%。

第 2 章　统计数据的收集与处理

一、填空题

1. 直接来源；间接来源
2. 间接来源数据
3. 普查；抽样调查
4. 经济性、时效性较高、适用面广、准确性较高
5. 概率抽样；非概率抽样
6. 非抽样误差；抽样误差

二、选择题

序号	1	2	3	4
答案	D	D	A	D

第 3 章　统计数据的整理与图形展示

一、填空题

1. 分类数据；顺序数据　2. 频数分布表　3. 频数　4. 频数分布表　5. 条形图

6. 帕累托图　7. 茎叶图　8. 茎；叶　9. 分组；未分组　10. 箱线图　11. 组距分组

12. 等距分组；不等距分组　13. 下限；上限　14. 组距　15. 不重不漏　16. 直方图

17. 时间序列数据　18. 二维散点图　19. 表头；行标题；列标题；数字资料

20. 表号；总标题；表中数据的单位

二、单项选择题

序号	1	2	3	4	5	6	7	8	9	10
答案	D	B	B	C	A	B	A	D	C	A

三、练习题

1.

(1) 顺序数据　(2) (3) 略

2～5 略

第 4 章　统计数据的指标度量

一、填空题

1．众数　　2．中位数　　3．四分位差　　4．对称分布　　5．0.4

6．偏态系数　　7．0.5　　8．0　　9．正值　　10．$Z=\frac{x_i-\bar{x}}{s}$

二、单项选择题

序号	1	2	3	4	5	6	7	8
答案	C	B	A	B	B	C	A	B
序号	9	10	11	12	13	14	15	
答案	B	D	A	D	B	C	A	

三、多项选择题

序号	1	2	3	4	5	6	7	8
答案	A B C	A B C	A B	A B C	A B	A B C	A B	A B

四、名词解释(略)

五、计算题

1．

首先将 25 个数据进行排序，排序结果如下：60、60、62、63、64、64、66、67、68、68、69、70、73、74、76、78、81、81、81、81、81、86、87、89、90

(1) 众数为 81

(2) 中位数：73

下四分位数：64+0.25*(66−64)＝64.5；上四分位数：81

(3) 平均数：73.56

方差：89.09

(4) 将这组数据输入到 Excel 中，则有：

众数的 Excel 的计算过程 MODE (60,60,…,90)

中位数的 Excel 的计算过程 MEDIAN(60,60,…，90)

下四分位数的 Excel 的计算过程 QUARTILE(array quart)，其中 array 为(60,60,…，90)，quart 为 1

上四分位数的 Excel 的计算过程 QUARTILE(array quart)，其中 array 为(60,60,…，90)，quart 为 3

平均数的 Excel 的计算过程 AVERAGE(60,60,…，90)

方差的 Excel 的计算过程 VAR(60,60,…，90)

2．

统计学成绩理想。(提示：计算两门成绩的标准分数)

3．

(1) 女童身高差异大。原因女童身高数据的离散系数大于男童身高数据的离散系数。

(2) 同(1)答案。原因：离散系数这个衡量指标本身就已消除了因计量单位不同或平均水平高低不等的影响

4．

$$\bar{x}=\frac{M_1f_1+M_2f_2+\cdots+M_kf_k}{f_1+f_2+\cdots+f_k}=\frac{750\times19+1\,250\times30+1\,750\times42+2\,250\times18+2\,750\times11}{120}=1\,633.33$$

$$s^2=\frac{\sum_{i=1}^{k}f_i(M_i-\bar{x})^2}{n-1}$$

$$=\frac{(750-1\,633)^2\times19+(1\,250-1\,633)^2\times30+(1\,750-1\,633)^2\times42+(2\,250-1\,633)^2\times18+(2\,750-1\,633)^2\times11}{120-1}$$

=339 125.8

5.

采用“离散系数”指标比较两种组装方法的离散程度

方法 1 的平均数为 166，标准差为 3.461 4，则离散系数为 47.96；方法 2 的平均数为 128，标准差为 1.912，则离散系数为 66.999 6，所以会选择方法进行组装，原因是平均值比方法 2 的平均数大，且离散系数比方法 2 的离散系数小，表明单位时间组装产品个数波动程度小。

第 5 章　参数估计

一、填空题

1. 统计量　2. 正态分布　3. 比例　4. 估计量　5. 估计值
6. 点估计　7. 区间估计　8. 置信区间，置信下限，置信上限
9. 置信水平　10. 无偏性，有效性，一致性　11. 一致性　12. 正态分布
13. 边际误差　14. 点估计值，边际误差　15. 窄

二、单项选择题

序号	1	2	3	4	5	6	7	8	9	10
答案	A	C	D	A	D	D	C	D	B	B
序号	11	12	13	14	15	16	17	18	19	20
答案	B	B	B	C	A	C	D	A	C	A
序号	21	22	23	24	25	26	27	28	29	30
答案	A	B	C	D	C	D	C	A	A	B

三、计算题

1.

(1) 81　(2) 1.2　(3) 2.352　(4) 78.648～83.352

2.

2/3

3.

根据：$\bar{X}\sim N(\mu,\frac{1}{n}\sigma^2)$ 有 $\bar{X}_{20}\sim N(10,500)$，$\bar{X}_{50}\sim N(10,200)$

4.

(1) 0.5　(2) 0.05

5.

(1) 样本比例的标准差公式为 $\sqrt{\frac{\pi(1-\pi)}{n}}$，所以分别为 0.04，0.0179，0.0126

(2) 随着样本容量的增大，样本比例的标准差越来越小

6. 提示 $\bar{x}\pm Z_{\alpha/2}\,s\big/\sqrt{n}$

7. 提示 $\bar{x}\pm t_{\alpha/2}(n-1)s\big/\sqrt{n}$

8. 提示 $\bar{x} \pm Z_{\alpha/2} s/\sqrt{n}$

9. 提示 $p \pm Z_{\alpha/2}\sqrt{\dfrac{p(1-p)}{n}}$

10. 提示 $\dfrac{(n-1)s^2}{\chi^2_{\alpha/2}(n-1)} \leqslant \sigma^2 \leqslant \dfrac{(n-1)s^2}{\chi^2_{1-\alpha/2}(n-1)}$

11. 提示 $n \geqslant \dfrac{(Z_{\alpha/2})^2\sigma^2}{E^2}$

12. 提示 $n \geqslant \dfrac{(Z_{\alpha/2})^2\pi(1-\pi)}{E^2}$

第 6 章　假设检验

一、单项选择题

序号	1	2	3	4	5	6	7	8	9	10	11	12	13
答案	A	B	A	C	A	B	D	B	C	D	A	C	B
序号	14	15	16	17	18	19	20	21	22	23	24	25	26
答案	A	D	B	A	A	B	A	B	C	A	D	A	B

二、简答题

1.

(1) 提出原假设 H_0 和备择假设 H_1。

(2) 构造检验的统计量，并计算其值。

(3) 根据给出的显著性水平 α，确定拒绝原假设 H_0 的区域。

(4) 统计决策。

2.

在样本容量不变的情况下，要减小 α 就会使 β 增大，而要增大 α 就会使 β 减小，这两类错误就像一个跷跷板。自然，人们希望犯两类错误的概率都尽可能小，但实际上很难做到两者错误发生概率同时减小。

3.

当总体方差已知，无论大小样本，假设检验统计量为 $Z=\dfrac{\bar{x}-\mu_0}{\sigma/\sqrt{n}} \sim N(0,1)$

当总体方差未知，大样本，假设检验统计量为 $Z=\dfrac{\bar{x}-\mu_0}{s/\sqrt{n}}$

当总体方差未知，小样本，假设检验统计量为 $t=\dfrac{\bar{x}-\mu_0}{s/\sqrt{n}} \sim t(n-1)$

4.

(1) H_0： $\pi=40\%$； H_1： $\pi \neq 40\%$

(2) $Z=\dfrac{p-\pi_0}{\sqrt{\dfrac{\pi_0(1-\pi_0)}{n}}} \sim N(0,1)$

即： $Z=\dfrac{p-40\%}{\sqrt{\dfrac{40\%(1-40\%)}{n}}} \sim N(0,1)$

(3) $(-\infty,-Z_{\alpha/2}) \cup (Z_{\alpha/2},+\infty)$，查表 $Z_{0.025}=1.96$，所以拒绝域为 $(-\infty,-1.96) \cup (1.96,+\infty)$。

(4) 作出决策。

三、判断分析题

1．

错误，正确的过程如下：

(1) 提出原假设和备择假设。

$$H_0:\sigma^2 \geqslant 0.001\,56; H_1:\sigma^2 < 0.001\,56$$

(2) 构造检验的统计量，并计算其值。

$$\chi^2=\frac{(n-1)S^2}{\sigma_0^2}=\frac{99\times 0.002\,11}{0.001\,56}=133.9$$

(3) $\alpha=0.05$，确定拒绝原假设的区域。

从备择假设中可以看出此处是左侧检验，所以拒绝域为$\left[0,\chi_{1-\alpha}{}^2(n-1)\right]$，其中$\chi_{0.95}{}^2(99)$查表得$\chi_{0.95}{}^2(99)=77.046\,3$，所以拒绝域为$(0,77.046\,3)$。

(4) 统计决策。

$\chi^2=133.9>\chi_{0.95}{}^2(99)=77.046\,3$，所以不拒绝原假设。

2．错误，正确的过程为：

(1) 提出原假设和备择假设。

$$H_0:\mu\leqslant 10; H_1:\mu>10$$

(2) 构造检验的统计量，并计算其值。

$$t=\frac{\bar{x}-\mu}{s/\sqrt{n}}=\frac{11-10}{5/\sqrt{25}}=1$$

(3) $\alpha=0.05$，确定拒绝原假设的区域。

备择假设中可以看出此处是右侧检验，所以拒绝域为$\left[t_\alpha(n-1),+\infty\right]$，其中$t_{0.05}(24)$查表得$t_{0.05}(24)=2.063\,9$，所以拒绝域为$(2.063\,9,+\infty)$

(4) 统计决策。

$t=1<t_{0.05}(24)=2.063\,9$，所以不拒绝原假设。

四、计算题

1．提示：$Z=\dfrac{\bar{x}-\mu_0}{s/\sqrt{n}}$

2．$Z=\dfrac{p-\pi_0}{\sqrt{\dfrac{\pi_0(1-\pi_0)}{n}}}\sim N(0,1)$

3．$Z=\dfrac{\bar{x}-\mu_0}{s/\sqrt{n}}$

4．$t=\dfrac{\bar{x}-\mu_0}{s/\sqrt{n}}\sim t(n-1)$

五、Excel 操作题(略)

第 7 章　方差分析

一、单项选择题

序号	1	2	3	4	5	6	7	8
答案	B	B	D	A	A	C	A	A
序号	9	10	11	12	13	14	15	
答案	A	A	A	B	B	A	A	

二、简答题

1．略

2．

(1) 每个总体都应服从正态分布。

(2) 方差齐性。

(3) 观测值是独立的。

3．

(1) 提出原假设 H_0 和备择假设 H_1

$H_0: \mu_1 = \mu_2 = \cdots = \mu_K$；　$H_1: \mu_1, \mu_2, \ldots, \mu_K$

(2) 构造检验统计量，并计算其值。

$$F = \frac{\text{SSA}/(k-1)}{\text{SSE}/(n-k)} = \frac{\text{MSA}}{\text{MSE}} \sim F(k-1, n-k)$$

(3) 根据给出的显著性水平 α，确定拒绝原假设的区域

$(F_\alpha(k-1, n-k), +\infty)$

(4) 统计决策。

当 $F > F_\alpha(k-1, n-k)$，拒绝原假设；当 $F \leqslant F_\alpha(k-1, n-k)$，不拒绝原假设。



4．

(1) 提出原假设和备择假设 $H_0: \mu_i = \mu_j$；　$H_1: \mu_i \neq \mu_j \ (i \neq j)$。

(2) 构造检验统计量，并计算其值 $\bar{x}_i - \bar{x}_j$。

(3) 根据给出的显著性水平 α 的数值，确定拒绝原假设的区域 $(-\infty, -\text{LSD}) \cup (\text{LSD}, +\infty)$。

(4) 作出决策。

三、计算题

1．

$\bar{x}_1 = 251, \bar{x}_2 = 255, \bar{x}_3 = 252, \bar{x}_4 = 252, \bar{x} = 253$

SSE＝696，SSA=75，$F = \dfrac{75/3}{696/26} = 0.933\ 9$，临界值为 4.636 7

4 个生产线的装填量无显著性差异。

2．

均值分别为 $\bar{x}_1 = 35.85, \bar{x}_2 = 31.633\ 3, \bar{x}_3 = 34.44, \bar{x} = 33.693\ 33$

SSE=74.975 33

SSA=46.854

$$F = \frac{46.854/2}{74.975\ 33/12} = 3.749\ 553$$

查表 $F_{0.01}(2,12)=6.926608$，所以不拒绝原假设，即 3 个路段对行车时间无显著性影响。

3.

(1)

差异源	SS	df	MS	F	F-crit
组间	420	2	210	1.4716	3.354 131
组内	3 836	27	142.07	—	—
总计	4256	29	—	—	—

(2) 3 个时间段的行车时间无显著性的差异。

四、Excel 操作题(略)

第 8 章　相关与一元回归分析

一、单项选择题

序号	1	2	3	4	5	6	7	8	9	10
答案	A	A	A	B	D	C	A	A	B	A
序号	11	12	13	14	15	16	17	18	19	20
答案	A	A	A	A	B	D	A	A	C	B

二、多项选择题

序号	1	2	3	4	5
答案	A B	A B C	A B	A B	A B

三、简答题

1.

(1) 进行相关分析。目的是判断因变量和自变量之间是否具有线性关系。

(2) 回归分析。如果第一步判断出变量之间存在线性关系，则要进行变量的回归分析。

(3) 经济预测和经济控制。这一步主要是利用第二步的回归分析，进行经济预测和经济控制。

2.

(1) 两变量的样本之间是否存在线性的关系；方法：散点图。

(2) 两变量的样本之间的关系强度如何；方法：相关系数。

(3) 样本所反映的变量之间的关系能否代表总体变量之间的关系。方法：相关系数检验。

3.

(1) 根据最小二乘法的定义可得。

$$L=\sum(y_i-\hat{y}_i)^2=\sum(y_i-\hat{\beta}_0-\hat{\beta}_1x_i)^2$$

(2) 求 $\min L$，一般情况，分别对上式中的未知数求一阶偏导，令其式子为 0。

$$\frac{\partial L}{\partial\hat{\beta}_0}=\sum -2\times(y_i-\hat{\beta}_0-\hat{\beta}_1x_i)=0$$

$$\frac{\partial L}{\partial\hat{\beta}_1}=\sum -2\times(y_i-\hat{\beta}_0-\hat{\beta}_1x_i)x_i=0$$

(3)

$$\hat{\beta}_0=\bar{y}-\hat{\beta}_1\bar{x}$$

$$\hat{\beta}_1=\frac{\sum x_i y_i-n\overline{xy}}{\sum x_i^2-n\overline{x}^2}$$

4.

1) 线性关系检验过程

(1) 提出原假设 H_0 和备择假设 H_1。

$H_0:\beta_1=0$　线性关系不显著

$H_1:\beta_1\neq 0$　线性关系显著

(2) 构造检验的统计量，并计算其值。

$$F=\frac{\mathrm{SSR}/1}{\mathrm{SSE}/(n-2)}\sim F(1,n-2)$$

(3) 根据给出的显著性水平 α，确定拒绝原假设的区域。

$$(F_\alpha(1,n-2),+\infty)$$

(4) 统计决策。

当 $F>F_\alpha(1,n-2)$ 时，检验统计量落在拒绝原假设区域内，所以拒绝原假设，两变量的线性关系显著。

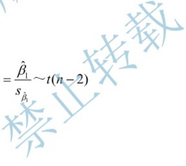
2) 回归系数的检验

(1) 回归系数 γ 的检验。

① 提出原假设 $0\leqslant\gamma\leqslant 1$ 和备择假设 $H_0:\rho=0,H_1:\rho\neq 0$。

$H_0:\beta_1=0$；$H_1:\beta_1\neq 0$

② 构造检验的统计量，并计算其值。



$$t=\frac{\hat{\beta}_1}{s_{\hat{\beta}_1}}\sim t(n-2)$$

式中，$s_{\hat{\beta}_1}=\sqrt{\frac{1}{\sum(x_i-\overline{x})^2}s_e^2}$

③ 根据给出的显著性水平 α，确定拒绝原假设的区域。

$(-\infty,-t_{\alpha/2}(n-2))\cup(t_{\alpha/2}(n-2),+\infty)$

④ 统计决策。

当 $|t|>t_{\alpha/2}(n-2)$ 时，检验统计量落在拒绝原假设区域内，所以拒绝原假设，回归系数 β_1 显著。

(2) 回归系数 β_0 的检验。

① 提出原假设 H_0 和备择假设 H_1。

$H_0:\beta_0=0$；$H_1:\beta_0\neq 0$

② 构造检验的统计量，并计算其值。

$$t=\frac{\hat{\beta}_0}{s_{\hat{\beta}_0}}\sim t(n-2)$$

③ 根据给出的显著性水平 α，确定拒绝原假设的区域。

$(-\infty,-t_{\alpha/2}(n-2))\cup(t_{\alpha/2}(n-2),+\infty)$

④ 统计决策。

当 $|t|>t_{\alpha/2}(n-2)$ 时，检验统计量落在拒绝原假设区域内，所以拒绝原假设，回归系数 β_0 显著。

四、计算题

1．略

2．

(1) 完成方差分析表和参数估计表。

方差分析表

变差来源	自由度	平方和	均方	*F*	*P*	*F*-crit
回归	1	1 744 705.4	1 744 705.4	194.4141	8.48119E-06	5.987 4
残差	6	53 845.02	8974．17	—	—	—
总计	7	1 798 550.42	—	—	—	—

参数估计表

	Cofficients	标准误差	*t* Stat	*P*
Intercep	5 050.5	867.783 5	5.82	0.000
X Variable	0.658	0.032 93	19.98	2.17E-09

(2) $\hat{y}_i = 5\,050.5 + 0.658x_i$ 。

(3) $\mathrm{R}^2 = \dfrac{SSR}{SST} = \dfrac{1\,744\,705.4}{1\,798\,550.42} = 0.97$ 。

(4) 写出线性关系检验的步骤。

① 提出原假设 H_0 和备择假设 H_1 。

$H_0 : \beta_1 = 0$　线性关系不显著

$H_1 : \beta_1 \neq 0$　线性关系显著



② 构造检验的统计量，并计算其值。

$F = \dfrac{\mathrm{SSR}/1}{\mathrm{SSE}/(n-2)} = 194.414\,1 \sim F(1, n-2)$

③ 根据给出的显著性水平 α ，确定拒绝原假设的区域。

$(F_\alpha(1,n-2),+\infty)$ 查表 $F_{0.05}(1,6) = 5.987\,4$

④ 统计决策。

当 $F > F_\alpha(1,n-2)$ 时，检验统计量落在拒绝原假设区域内，所以拒绝原假设，两变量的线性关系显著。

(5) 写出自变量回归系数的检验步骤。

① 提出原假设 $0 \leqslant \gamma \leqslant 1$ 和备择假设 $H_0 : \rho = 0, H_1 : \rho \neq 0$ 。

$H_0 : \beta_1 = 0$ ； $H_1 : \beta_1 \neq 0$

② 构造检验的统计量，并计算其值。

$$t = \frac{\hat{\beta}_1}{s_{\hat{\beta}_1}} = 19.98 \sim t(n-2)$$

③ 根据给出的显著性水平 α ，确定拒绝原假设的区域。

$(-\infty, -t_{\alpha/2}(n-2)) \cup (t_{\alpha/2}(n-2), +\infty)$ 查表 $t_{0.025}(6) = 2.968\,7$

④ 统计决策。

当 $|t| > t_{\alpha/2}(n-2)$ 时，检验统计量落在拒绝原假设区域内，所以拒绝原假设，回归系数 β_1 显著。

(6) 略。

北京大学出版社本科财经管理类实用规划教材(已出版)

财务会计类

序号	书　名	标 准 书 号	主　编	定价	序号	书　名	标 准 书 号	主　编	定价
1	基础会计	7-301-24366-4	孟　铁	35.00	22	中级财务会计	7-301-23772-4	吴海燕	49.00
2	基础会计(第 2 版)	7-301-17478-4	李秀莲	38.00	23	中级财务会计习题集	7-301-25756-2	吴海燕	39.00
3	基础会计实验与习题	7-301-22387-1	左　旭	30.00	24	高级财务会计	7-81117-545-5	程明娥	46.00
4	基础会计学	7-301-19403-4	窦亚芹	33.00	25	高级财务会计	7-5655-0061-9	王奇杰	44.00
5	基础会计学学习指导与习题集	7-301-16309-2	裴　玉	28.00	26	企业财务会计模拟实习教程	7-5655-0404-4	董晓平	25.00
6	基础会计	7-301-23109-8	田凤彩	39.00	27	成本会计学	7-301-19400-3	杨尚军	38.00
7	基础会计学	7-301-16308-5	晋晓琴	39.00	28	成本会计学	7-5655-0482-2	张红漫	30.00
8	信息化会计实务	7-301-24730-3	杜天宇	35.00	29	成本会计学	7-301-20473-3	刘建中	38.00
9	会计学原理习题与实验(第 2 版)	7-301-19449-2	王保忠	30.00	30	税法与税务会计实用教程(第 2 版)	7-301-21422-0	张巧良	45.00
10	会计学原理(第 3 版)	7-301-26239-9	刘爱香	35.00	31	初级财务管理	7-301-20019-3	胡淑姣	42.00
11	会计学原理	7-301-24872-0	郭松克	38.00	32	财务会计学	7-301-23190-6	李柏生	39.00
12	会计学原理与实务(第 2 版)	7-301-18653-4	周慧滨	33.00	33	财务管理学实用教程(第 2 版)	7-301-21060-4	骆永菊	42.00
13	初级财务会计模拟实训教程	7-301-23864-6	王明珠	25.00	34	财务管理理论与实务(第 2 版)	7-301-20407-8	张思强	42.00
14	初级会计学习题集	7-301-25671-8	张兴东	28.00	35	财务管理理论与实务	7-301-20042-1	成　兵	40.00
15	会计规范专题(第 2 版)	7-301-23797-7	谢万健	42.00	36	财务管理学	7-301-21887-7	陈　玮	44.00
16	会计综合实训模拟教程	7-301-20730-7	章洁倩	33.00	37	公司财务管理	7-301-21423-7	胡振兴	48.00
17	预算会计	7-301-22203-4	王筱萍	32.00	38	财务分析学	7-301-20275-3	张献英	30.00
18	会计电算化	7-301-23565-2	童　伟	49.00	39	审计学	7-301-20906-6	赵晓波	38.00
19	政府与非营利组织会计	7-301-21504-3	张　丹	40.00	40	审计理论与实务	7-81117-955-2	宋传联	36.00
20	管理会计	7-81117-943-9	齐殿伟	27.00	41	现代审计学	7-301-25365-6	杨　茁	39.00
21	管理会计	7-301-21057-4	彤芳珍	36.00					

管理类

序号	书　名	标 准 书 号	主　编	定价	序号	书　名	标 准 书 号	主　编	定价
1	管理学	7-301-17452-4	王慧娟	42.00	14	统计学	7-301-24750-1	李付梅	39.00
2	管理学	7-301-21167-0	陈文汉	35.00	15	统计学	7-301-25180-5	邓正林	42.00
3	管理学	7-301-23023-7	申文青	40.00	16	统计学(第 2 版)	7-301-23854-7	阮红伟	35.00
4	管理学原理	7-301-22980-4	陈　阳	48.00	17	应用统计学(第 2 版)	7-301-19295-5	王淑芬	48.00
5	管理学原理	7-5655-0078-7	尹少华	42.00	18	统计学理论与实务	7-301-24467-8	王雪秋	30.00
6	管理学原理	7-301-21178-6	雷金荣	39.00	19	统计学实验教程	7-301-22450-2	裘雨明	24.00
7	管理学原理与实务(第 2 版)	7-301-18536-0	陈嘉莉	42.00	20	管理运筹学(第 2 版)	7-301-19351-8	关文忠	39.00
8	现代企业管理理论与应用(第 2 版)	7-301-21603-3	邸彦彪	38.00	21	企业经营 ERP 沙盘模拟教程(第 2 版)	7-301-26163-7	董红杰	45.00
9	管理学实用教程	7-301-21059-8	高爱霞	42.00	22	项目管理	7-301-21448-0	程　敏	39.00
10	新编现代企业管理	7-301-21121-2	姚丽娜	48.00	23	项目管理	7-301-24823-2	康　乐	39.00
11	统计学原理(第 2 版)	7-301-25114-0	刘晓利	36.00	24	公司治理学	7-301-22568-4	蔡　锐	35.00
12	统计学原理	7-301-21061-4	韩　宇	38.00	25	现场管理	7-301-21528-9	陈国华	38.00
13	统计学原理与实务	7-5655-0505-8	徐静霞	40.00					

市场营销类

序号	书　名	标 准 书 号	主　编	定价	序号	书　名	标 准 书 号	主　编	定价
1	市场营销学	7-301-21056-7	马慧敏	42.00	6	市场营销学(第 2 版)	7-301-19855-1	陈　阳	45.00
2	市场营销学：理论、案例与实训	7-301-21165-6	袁连升	42.00	7	市场营销学	7-301-21166-3	杨　楠	40.00
3	市场营销学实用教程(第 2 版)	7-301-24958-1	林小兰	48.00	8	市场营销理论与实务(第 2 版)	7-301-20628-7	那　薇	40.00
4	市场营销学(第 2 版)	7-301-24328-2	王槐林	39.00	9	国际市场营销学	7-301-21888-4	董　飞	45.00
5	营销策划	7-301-23204-0	杨　楠	42.00	10	营销策划	7-301-26027-2	张　娟	38.00

序号	书　名	标 准 书 号	主　编	定价	序号	书　名	标 准 书 号	主　编	定价
11	市场营销策划	7-301-23384-9	杨　勇	40.00	16	客户关系管理理论与实务	7-301-23911-7	徐　伟	40.00
12	广告策划与管理：原理、案例与项目实训	7-301-23827-1	杨佐飞	48.00	17	社交礼仪	7-301-23418-1	李　霞	29.00
13	现代推销与谈判实用教程	7-301-25695-4	凌奎才	48.00	18	商务谈判(第 2 版)	7-301-20048-3	郭秀君	49.00
14	消费者行为学	7-5655-0057-2	肖　立	37.00	19	消费心理学(第 2 版)	7-301-25983-2	臧良运	40.00
15	客户关系管理实务	7-301-09956-8	周贺来	44.00					

工商管理类

序号	书　名	标 准 书 号	主　编	定价	序号	书　名	标 准 书 号	主　编	定价
1	企业文化理论与实务(第 2 版)	7-301-24445-6	王水嫩	35.00	10	创业基础：理论应用与实训实练	7-301-24465-4	郭占元	38.00
2	企业战略管理实用教程	7-81117-853-1	刘松先	35.00	11	公共关系学实用教程(第 2 版)	7-301-25557-5	周　华	42.00
3	企业战略管理	7-301-23419-8	顾　桥	46.00	12	公共关系学实用教程	7-301-17472-2	任焕琴	42.00
4	生产运作管理(第 3 版)	7-301-24502-6	李全喜	54.00	13	公共关系理论与实务	7-5655-0155-5	李泓欣	45.00
5	运作管理	7-5655-0472-3	周建亨	25.00	14	东方哲学与企业文化	7-5655-0433-4	刘峰涛	34.00
6	运营管理实验教程	7-301-25879-8	冯根尧	24.00	15	跨国公司管理	7-5038-4999-2	冯雷鸣	28.00
7	组织行为学实用教程	7-301-20466-5	冀　鸿	32.00	16	企业战略管理	7-5655-0370-2	代海涛	36.00
8	质量管理(第 2 版)	7-301-24632-0	陈国华	39.00	17	跨文化管理	7-301-20027-8	晏　雄	35.00
9	创业学	7-301-15915-6	刘沁玲	38.00					

人力资源管理类

序号	书　名	标 准 书 号	主　编	定价	序号	书　名	标 准 书 号	主　编	定价
1	人力资源管理(第 2 版)	7-301-19098-2	颜爱民	60.00	5	员工招聘	7-301-20089-6	王　挺	30.00
2	人力资源管理实用教程(第 2 版)	7-301-20281-4	吴宝华	45.00	6	人力资源管理：理论、实务与艺术	7-5655-0193-7	李长江	48.00
3	人力资源管理原理与实务(第 2 版)	7-301-25511-7	邹　华	32.00	7	人力资源管理实验教程	7-301-23078-7	畅铁民	40.00
4	人力资源管理教程	7-301-24615-3	夏兆敢	36.00					

服务管理类

序号	书　名	书号	编著者	定价	序号	书　名	书号	编著者	定价
1	会展服务管理	7-301-16661-1	许传宏	36.00	4	服务性企业战略管理	7-301-20043-8	黄其新	28.00
2	非营利组织管理	7-301-20726-0	王智慧	33.00	5	现代服务业管理原理、方法与案例	7-301-17817-1	马　勇	49.00
3	服务营销	7-301-21889-1	熊　凯	45.00					

如您需要更多教学资源如电子课件、电子样章、习题答案等，请登录北京大学出版社第六事业部官网 www.pup6.cn 搜索下载。

如您需要浏览更多专业教材，请扫下面的二维码，关注北京大学出版社第六事业部官方微信(微信号：pup6book)，随时查询专业教材、浏览教材目录、内容简介等信息，并可在线申请纸质样书用于教学。

感谢您使用我们的教材，欢迎您随时与我们联系，我们将及时做好全方位的服务。联系方式：010-62750667，wangxc02@163.com，pup_6@163.com，lihu80@163.com，欢迎来电来信。客户服务 QQ 号：1292552107，欢迎随时咨询。